MARKOV RANDOM FIELD IMAGE MODELLING

By

Michael Mc Grath

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
AT THE
UNIVERSITY OF CAPE TOWN
CAPE TOWN, SOUTH AFRICA
JUNE 2003

Declaration

I, Michael Mc Grath, declare that this dissertation is my own work. It is being submitted for the degree of Master of Science in Engineering at the University of Cape Town and it has not been submitted before for any degree or examination, at any other university.

signature removed

Michael Mc Grath

Abstract

This work investigated some of the consequences of using *a priori* information in image processing using computer tomography (CT) as an example. Prior information is information about the solution that is known apart from measurement data. This information can be represented as a probability distribution. In order to define a probability density distribution in high dimensional problems like those found in image processing it becomes necessary to adopt some form of parametric model for the distribution. Markov random fields (MRFs) provide just such a vehicle for modelling the *a priori* distribution of labels found in images.

In particular, this work investigated the suitability of MRF models for modelling *a priori* information about the distribution of attenuation coefficients found in CT scans. This involved selecting different models and fitting them to sample images of CT scans. These MRF models were then used in a number of experiments and were found to lead to more accurate tomographic reconstructions.

In the experiments maximum *a posteriori* (MAP) estimation using MRFs to model the *a priori* distribution was found to outperform maximum likelihood (ML) estimation which does not use prior information. The experiments included cases where the angular range was less than 180 degrees (limited angle tomography) and cases were the angular range was sparse (sparse angle tomography).

Acknowledgements

Warm thanks to my supervisor Professor Gerhard DeJager who supported me even when progress was slow. I would also like to thank AMI and the NRF for financial support.

I would also like to thank my friends in the Digital Image Processing group and those from AMI for making my studies an enjoyable experience.

Finally, a word of thanks must go to my family and to Jesus Christ, thanks for your love.

Cape Town,
December, 2002

Michael McGrath

Table of Contents

Chapter 1

Introduction

Computer tomography allows internal anatomical detail of a patient to be examined with minimal danger to the patient. For this reason computer tomography (CT) has revolutionized medical practice since the pioneering work of Allan Cormack and Godfrey Hounsfield who both received the Nobel prize in Physiology or Medicine in 1979 [25].

Since then computer tomography has reached a mature state of development with commercial machines producing good quality reconstructions in reasonable time due to efficient reconstruction algorithms. These algorithms fail when their sample requirements are not met. Conditions under which these algorithms fail include cases where the projection data is only available over a limited angular range, cases where projection data are only available at a few projection angles, and cases where the data measurements are noisy.

In cases where the available data is insufficient to specify a unique solution, the problem is said to be ill-posed. If methods using all statistical information about the measurement process fail to produce sufficiently good results one has the choice of either giving up or of bringing other knowledge to bear on the problem. This type of knowledge is called *a priori* knowledge and is knowledge about the solution that does not come from the measurement data. An example of *a priori* knowledge in computer tomography is that X-ray attenuation

coefficients cannot be negative as this would mean that more X-ray photons were leaving a region than were entering it. This information could be used to improve an estimate of the attenuation coefficients of an object. In fact, 'It is is fundamental rule of estimation theory that the use of prior knowledge will lead to a more accurate estimator' [34].

All knowledge about possible estimates can be represented as a probability distribution that assigns a probability to each possible solution. This distribution is called the *a priori* probability distribution.

For problems of high dimension, like computer tomography, the configuration space of possible solutions is very large, making the direct definition of the probability distribution unfeasible. In order to define a probability distribution in high dimensional problems like this, it becomes necessary to adopt some form of parametric model for the distribution. Markov Random Fields (MRFs) provide just such a vehicle for modelling the *a priori* distribution of images.

The aim of this work has been to investigate some of the consequences of using *a priori* information in image processing and computer tomography. In particular, it investigated the suitability of Markov random field models for modelling *a priori* information about the distribution of attenuation coefficients found in CT scans. This involved selecting different models and fitting them to sample images. A secondary goal was to use these models to help solve some image processing problems and determine whether their use led to improved results over methods that do not take *a priori* information into account.

"Since its beginnings, computer vision research has been evolving from heuristic design of algorithms to systematic investigation of approaches."

—Stan Z. Li [1]

1.1 Markov Random Fields

Markov random field theory holds the promise of providing a systematic approach to the analysis of images in the framework of Bayesian probability theory. Markov random fields (MRFs) model the statistical properties of images. This allows a host of statistical tools and approaches to be turned to solving so called *ill-posed problems* in which the measured data does not specify a unique solution.

This chapter introduces a number of concepts needed to understand Markov random fields and how they may be used for modelling images. Defining a probability density distribution for an image requires that a probability mass be assigned to each possible configuration of labels or intensities in an image. As this configuration space is very large and cannot be calculated directly, parametric methods are needed. MRFs can be used as parametric models for the probability distribution of intensity levels in an image. In more abstract terms this can be seen as modelling the distribution of labels on a set of sites.

1.2 Sites and Labels

A Markov random field is defined on a set of sites. The sites may be regularly spaced on a lattice or irregularly spaced. Regularly spaced sites are suitable for modelling pixel

[1] From page XI of his book *Markov Random Field Modeling in Image Analysis* [38]

intensity levels in images and will be used throughout this work. Irregularly spaced sites are useful for high level vision problems in which features have been extracted from the image. Irregularly spaced sites are usually referred to in the statistical literature as point processes rather than Markov random fields [43]. Let $\mathcal{S}$ be a set of m discrete sites

$$\mathcal{S} = \{1, \ldots, m\} \tag{1.1}$$

in which $1, \ldots, m$ are indices. A set of sites on a square $n \times n$ lattice can also be written as $\mathcal{S} = \{(i, j) | 1 \leq i, j \leq n\}$.

Each site has a label associated with it. The set of possible labels may be continuous or discrete. The adoption of either a continuous or a discrete label set is one of the first decisions that need to be made as this determines the nature of the solution space. If the label set is continuous, the probability distribution used to model the problem must also be continuous in which case it is known as a probability density function. If the label set is discrete, the probability distribution used to model the problem must also be discrete and is called a probability mass function. For now, a set $\mathcal{L}$ of M discrete labels will be adopted such that

$$\mathcal{L} = \{l_1, ..., l_M\}. \tag{1.2}$$

The labelling for a set of sites, $\mathcal{S}$, will be denoted by

$$f = \{f_1, \ldots, f_m\} \tag{1.3}$$

where f_i is the label at site i. The set of all possible configurations is called F. The size of the configuration space F is given by M^m where M is the number of candidate labels for each site and m is the number of sites on the lattice. Many problems in machine vision can be cast into this form where the problem is to estimate the best labelling for a set of sites. For the example used in this chapter each site will be assigned one of four possible

labels such that $\mathcal{L} = \{l_1, l_2, l_3, l_4\}$. In this case the labels are unordered. This means that a statement like *label* l_4 *is greater than label* l_1 is meaningless. Unordered labels arise from classification problems where the image is divided up into a number of regions. Labels used to represent image intensities are more naturally treated as ordered. Examples of the use of ordered labels include image restoration [18] [52], surface reconstruction [16] and image reconstruction in computer tomography [2][15].

1.3 The Markov Property

The defining characteristic of MRFs is that the interaction between labels is limited to a local region. This region is called the neighbourhood of a site. The sites of a Markov random field on a lattice $\mathcal{S}$ are related to each other via a neighbourhood system, $\mathcal{N}$, such that

$$\mathcal{N} = \{\mathcal{N}_i | \forall i \in \mathcal{S}\} \tag{1.4}$$

where $\mathcal{N}_i$ is the set of sites neighbouring site i. A site cannot be a neighbour to itself. Figures 1.1 and 1.2 show the neighbourhoods for 4 and 8 neighbourhood models. The shaded square represents the site of interest and the white squares represent the neighbouring sites. The figures also show how the neighbourhood can be broken up into a number of cliques. A clique determines the arguments for the potential functions which define different Markov random field models. A clique for a site i must include that site as one of its members and may contain other sites in the neighbourhood of the site i. The concept of a neighbourhood system will be expanded upon in the next section.

A random process is said to be Markov if the following condition holds. The conditional probability function for the label at a site i given the labels of all other sites on $\mathcal{S}$ is equal

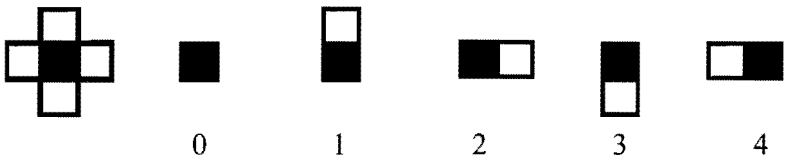


Figure 1.1: 1st order or 4 neighbourhood system and its division into cliques. The shaded squares represents the site of interest and the white squares represent the neighbouring sites.

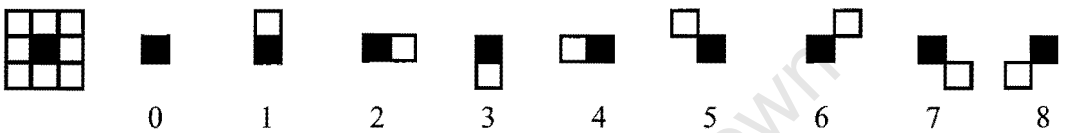


Figure 1.2: 2nd order 8 neighbourhood system and its division into cliques.

to the conditional probability for that label given only the labels in the neighbourhood of site i. Following the notation of Li's book [38], this can be written as

$$P(f_i|f_{S-\{i\}}) = P(f_i|f_{\mathcal{N}_i}). \tag{1.5}$$

Equation 1.5 does not mean that the labels of sites not in each others neighbourhood are independent, but rather that all information about the distribution at a site is given by its neighbours and no more information can be gained by considering sites outside of that sites neighbourhood. In other words, correlations may extend far beyond the local neighbourhood of a site [6].

The conditional distribution of a site gives the probability of possible labels at that site given the labels at neighbouring sites. It is difficult to specify a Markov random field by its conditional probability structure as there are highly restrictive consistency conditions [4]. Fortunately Gibbs distributions provide a way to specify a Markov random field by its joint probability distribution. The joint probability assigns a probability to each possible

configuration f on the lattice $\mathcal{S}$. It is the joint probability that is required for the maximum *a posteriori* estimation algorithm described in Chapter 6.

1.4 The Gibbs Distribution

Markov random fields and Gibbs distributions are equivalent. The Gibbs distribution of a Markov random field is just the joint probability of that Markov random field.

Let $P(f)$ be a Gibbs distribution on a lattice $\mathcal{S}$. Then $P(f)$ has a form given by

$$P(f) = Z^{-1} \times e^{-\frac{1}{T}U(f)} \tag{1.6}$$

where

$$Z = \sum_{f \in F} e^{-\frac{1}{T}U(f)} \tag{1.7}$$

is a normalizing constant called the partition function. Calculating the partition function exactly involves normalizing over all possible configurations which is computationally prohibitive for even moderately sized images as the number of possible configurations is given by M^m where M is the number of labels for each site and m is the number of sites on the lattice. The term Z is sometimes called the free energy of the system.

The energy function $U(f)$ in Equation 1.6 is the sum of clique potential functions, $V_c(f)$, over all cliques $\mathcal{C}$ on the lattice $\mathcal{S}$ as given by Equation 1.8. Configurations with higher energy have less probability of occurring.

$$U(f) = \sum_{c \in \mathcal{C}} V_c(f) \tag{1.8}$$

The energy $U(f)$ and the clique potential functions $V_c(f)$ should be positive for all possible label configurations, to enable correct normalization of Equation 1.6. This positivity

constraint can be enforced on clique potential functions by subtracting the minimum value of the potential function over the domain $\mathcal{L}$ from the potential function as shown in Equation 1.9. This is done for all clique potential functions except for the uniform prior defined in Equation 3.19 which is defined to be positive for all possible labels.

$$V_c(f) \leftarrow V_c(f) - \min_{l_i \in \mathcal{L}} V_c(l_i) \tag{1.9}$$

The order of a clique is given by the number of sites in the clique. A first order clique potential is thus a function of the label at one site. A second order clique potential is a function of the labels at two sites and is also the lowest order clique potential to convey contextual information or to model dependence between the labels at neighbouring sites.

The term T in Equation 1.6 is a scalar that represents temperature in physical systems and will be referred to as the temperature here. As the value of T is increased the distribution approaches a uniform distribution, for which each configuration has the same probability. Similarly, as the temperature is lowered the distribution becomes more peaked with the probability mass concentrating at the most likely configurations. The temperature term T is prominent in the simulated annealing optimization algorithm where the search strategy involves sampling the same distribution at different temperatures [45]. The simulated annealing algorithm will be further discussed in Chapter 2.

The following proof that a Gibbs distribution is equivalent to a Markov random field is taken from Li [38]. Consider the conditional probability for the label at a site i given the labels at all other sites on $\mathcal{S}$

$$P(f_i | f_{\mathcal{S}-\{i\}}) = \frac{P(f_i, f_{\mathcal{S}-\{i\}})}{P(f_{\mathcal{S}-\{i\}})} = \frac{P(f)}{\sum_{f'_i \in \mathcal{L}} P(f')} \tag{1.10}$$

where $f' = \{f_1, \ldots, f_{i-1}, f'_i, \ldots, f_m\}$ is any configuration which agrees with f at all sites except possibly at i. The notation $\mathcal{S} - \{i\}$ indicates the set of all sites on the lattice $\mathcal{S}$

excluding site i. Writing out $P(f) = Z^{-1} \times e^{-\sum_{c \in \mathcal{C}} V_c(f)}$ using Equation 1.6 and 1.8 gives

$$P(f_i|f_{S-\{i\}}) = \frac{e^{-\sum_{c \in \mathcal{C}} V_c(f)}}{\sum_{f_i'} e^{-\sum_{c \in \mathcal{C}} V_c(f')}}. \tag{1.11}$$

Divide $\mathcal{C}$ into two sets $\mathcal{A}$ and $\mathcal{B}$ with $\mathcal{A}$ consisting of cliques containing site i and $\mathcal{B}$ consisting of cliques not containing site i. Then Equation 1.11 can be written as

$$P(f_i|f_{S-\{i\}}) = \frac{[e^{-\sum_{c \in \mathcal{A}} V_c(f)}][e^{-\sum_{c \in \mathcal{B}} V_c(f)}]}{\sum_{f_i'}\{[e^{-\sum_{c \in \mathcal{A}} V_c(f')}][e^{-\sum_{c \in \mathcal{B}} V_c(f')}]\}}. \tag{1.12}$$

Because $V_c(f) = V_c(f')$ for any clique c that does not contain site i, $e^{-\sum_{c \in \mathcal{B}} V_c(f)}$ cancels from the numerator and denominator. Therefore, this probability depends only on the potentials of the cliques containing site i.

$$P(f_i|f_{S-\{i\}}) = \frac{e^{-\sum_{c \in \mathcal{A}} V_c(f)}}{\sum_{f_i'} e^{-\sum_{c \in \mathcal{A}} V_c(f')}} \tag{1.13}$$

This proves that a Gibbs random field is a Markov random field where the neighbourhood of i is determined by those clique potential functions that include site i. Thus if a site i is a neighbour of site j, site j is also a neighbour of site i. This can be written as, if $i \in \mathcal{N}_j$ then $j \in \mathcal{N}_i$.

The numerator of Equation 1.13 is the potential of the configuration at the site while the denominator is the normalizing factor taken over all possible labels for that site. Equation 1.13 also tells us how to calculate the conditional probability densities of the equivalent Markov random field from a Gibbs distribution.

1.5 The Multi-level Logistic Model

This section uses the multi-level logistic (MLL) model as an example of a MRF model [26]. The primary use of the multi-level logistic model is for modelling the distribution of

regions although it can also be used to model simple textures. Samples from MLL models are shown in Figure 1.3.

Different MRFs are realized by the choice of potential functions and the neighbourhoods over which they act. The potential functions for the MLL model can be defined as follows. The potential for each site is the sum of the contributions from a single site clique and those pairwise cliques that involve the site. The potential of a single site clique is a function of the label at that site.

$$V_c(f_i) = \begin{cases} \alpha_1 & \text{if } f_i = l_1 \\ \alpha_2 & \text{if } f_i = l_2 \\ \alpha_3 & \text{if } f_i = l_3 \\ \alpha_4 & \text{if } f_i = l_4 \end{cases} \tag{1.14}$$

The potential of single site cliques in a Markov random field should be related to the relative frequency or probability of each label. Assuming the following values for the probability of each label,

$$\begin{aligned} P(f_i = l_1) &= 0.1 \\ P(f_i = l_2) &= 0.2 \\ P(f_i = l_3) &= 0.4 \\ P(f_i = l_4) &= 0.3, \end{aligned} \tag{1.15}$$

the probability for each label can be written in the form of a Gibbs potential function.

$$\begin{aligned} P(f_i = l_1) &= e^{\ln(0.1)} \\ P(f_i = l_2) &= e^{\ln(0.2)} \\ P(f_i = l_3) &= e^{\ln(0.4)} \\ P(f_i = l_4) &= e^{\ln(0.3)} \end{aligned} \tag{1.16}$$

Thus for a site i, with the probability for label l_k given by $P(l_k)$ the $1st$ order clique potential function is given by $V_c(f_i = l_k) = -\ln(P(l_k))$. This shows that the probability of every label must be greater than zero for the Gibbs distribution to be defined as $\ln(0)$ is not

defined. If information about the relative frequencies of labels is not available a uniform distribution should be used where the probability of each label is the same. This choice is motivated by the principle of maximum entropy that states that when information about a distribution is incomplete the distribution with maximum entropy that agrees with the incomplete data should be chosen [10],[31]. When no data is available an uninformative or uniform distribution should be chosen that assigns the same probability to each possible configuration. Entropy is a measure of the amount of uncertainty in a probability distribution [31]. A uniform distribution is said to have maximum entropy and a distribution in which one event occurs with certainty is a minimum entropy distribution.

Single site cliques can only convey information on the relative frequency of different labels and cannot convey contextual information. To convey contextual information, cliques with two or more sites are needed.

When defining models the clique potential function for a clique will be given as the sum of all the clique potential functions on that clique. If conditional probabilities need to be calculated this notation is more natural, although care must be taken not to double count pairwise clique potentials when calculating the joint probability of the random field. This notation requires another constraint on the definition of MRF models, that there must be symmetry around the site being considered. Thus for a 1st order neighbourhood system only two pairwise clique potential functions need to be defined. One for the vertically aligned cliques and one for the horizontally aligned cliques. In terms of Figure 1.1 the potential functions for cliques 1 and 3 and cliques 2 and 4 must be the same. This notation can only be used for homogenous Markov random fields for which clique potential functions do not change with the position of a site on the lattice $\mathcal{S}$.

The potential function for pairwise cliques in the MLL model consisting of the site i

and one of its neighbours can be defined as follows.

$$V_C(f_i, f_i') = \begin{cases} \beta_c & \text{if sites on clique } \{i, i'\} \text{ have the same label} \\ -\beta_c & \text{otherwise} \end{cases} \tag{1.17}$$

Although this is the form in which the MLL potential function is usually defined it does introduce negative energy components and can be restated as

$$V_C(f_i, f_i') = \begin{cases} 2\beta_c & \text{if sites on clique } \{i, i'\} \text{ have the same label} \\ 0 & \text{otherwise.} \end{cases} \tag{1.18}$$

The form of the MLL model has now been defined. By changing the value of the α and β parameters, different distributions can be modelled. Figure 1.3 shows samples taken from MLL distributions for different parameter values. A uniform distribution was used for the $1st$ order cliques while all the second order cliques, in the 8 neighbourhood model used, share the same potential function defined by the parameter β. The images were simulated using one hundred iterations of a Metropolis sampler [44]. Image (a) and (b) were initialized from a constant flat image while (c) and (d) were initialized from random independent samples. This was done because for β larger than 0.4 the Markov random field is close to freezing and therefore strongly favours uniform images. If the image were initialized using a uniform image, the model would not be able to escape from this low energy configuration.

The images in Figure 1.3 can be interpreted as being generated by the same distribution at different temperatures. This is because the energy of each model is linearly related to the others. If image (d) is nominally assigned the temperature $T = 1$, then the temperature of image (a) is $T = 6$, the temperature of image (b) is $T = 3$ and the temperature of image (c) is given by $T = 1.5$.

Boundary sites may be dealt with in a number of ways. The simplest approach is to hold boundary sites constant as then all sites have the same clique configuration. Another

approach is to adopt toroidal periodicity where the lattice is wrapped into the shape of a doughnut [42]. The approach adopted in this work was to define the energy function $U(f)$ at the boundary sites using only those cliques that were defined on the lattice. The effect of this is that the sites on the boundary of the lattice tend to have a larger variance as their interaction with the lattice is weaker than for interior sites.

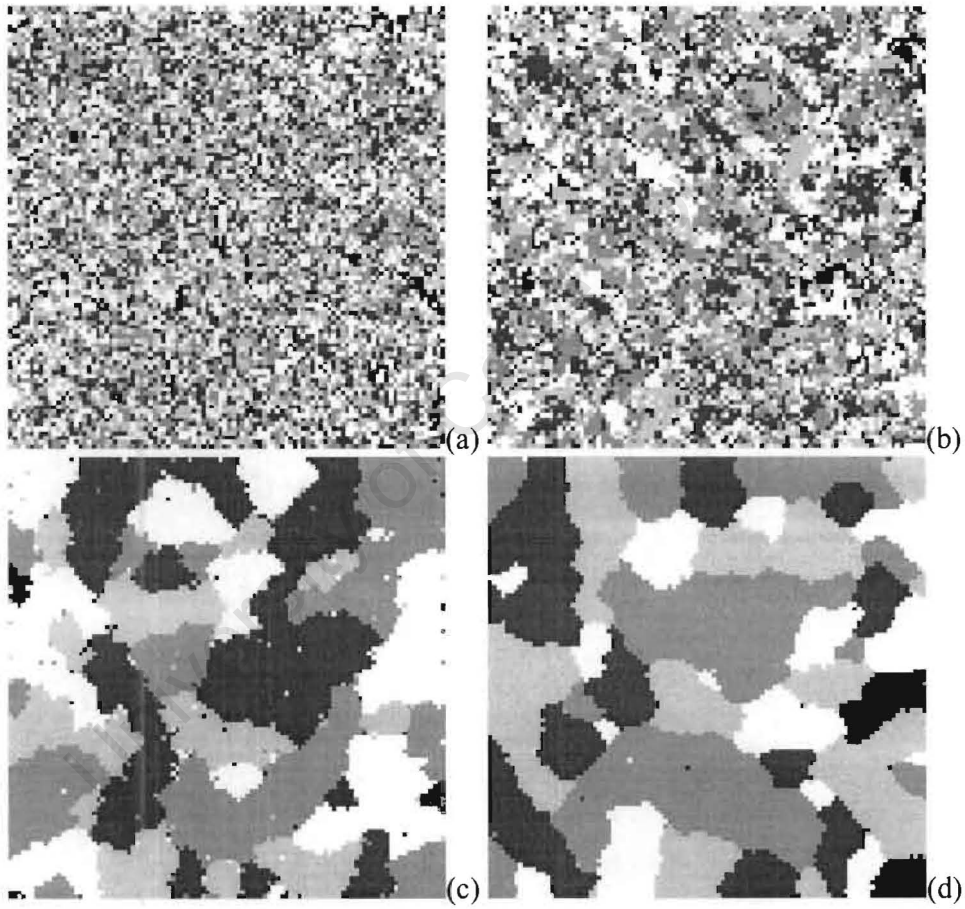


Figure 1.3: Sample images from the MLL model for different values of β, (a) $\beta = 0.1$ (b) $\beta = 0.2$ (c) $\beta = 0.4$ (d) $\beta = 0.6$

1.6 Some Underlying Assumptions

The underlying assumption of using Markov random fields is that an image can be treated as a sample from a random process. The validity of this assumption is not obvious for many images. In practice there are often statistical relationships between labels that can be modelled. Even for complex images like Figure 1.4 (a) it may be reasonable to model regions like those in Figure 1.4 (b) using MRFs. This assumption must be made in order to use sample images to train MRF models.

It is not strictly necessary to assume that an image can be treated as a sample from a random process as the role of the prior distribution is to represent our incomplete knowledge about the parameters of interest. It is not necessary that these parameters be samples from a random process. The prior distribution need not represent any physical property of the parameters, but only the state our knowledge about the parameters [31].

For example, 'To assign equal probabilities to two events is not in any way an assertion that they must occur equally often in any "random experiment"' [31]. Rather it is a way to show uncertainty or lack of knowledge about the events.

Rejecting the assumption that an image can be treated as a sample from a random process leaves one with the thorny problem of how to define the *a priori* distribution without recourse to training images and so is not done here.

An important assumption that is often made when using Markov random fields is that of homogeneity. This implies that the model does not change with position on the lattice. This assumption is important as is allows inferences to be made about the model by greatly reducing the dimensionality of the model.

The validity of this assumption is not obvious for many images. For complex images like Figure 1.4 (a) it may be necessary to model different regions like those in Figure 1.4

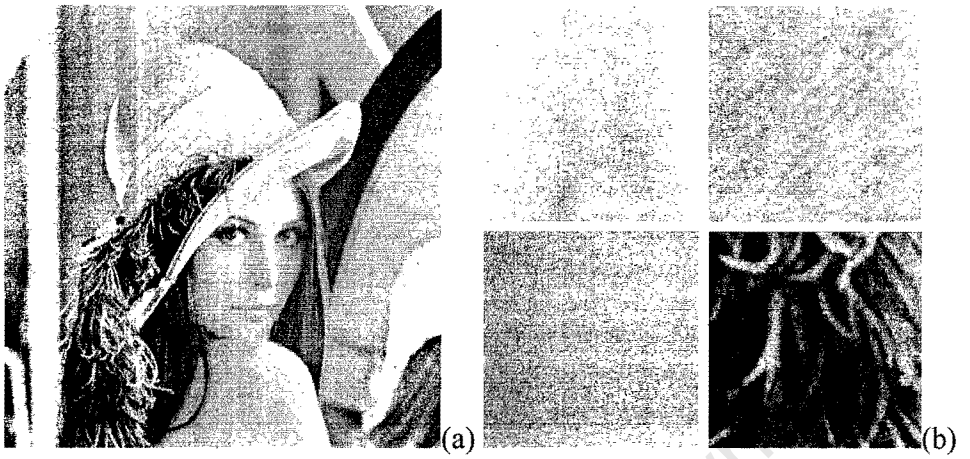


Figure 1.4: A complex image for which the assumption of homogeneity may not be valid. Image (b) shows details from image (a).

(b) using different MRF models.

The assumptions that a set of images can be treated as being homogeneous over their extent and that a set of images can be treated as being sampled from a random process becomes more reasonable when the modality for gathering image data does not change and the scale and subject matter of the images are similar. For example, a set of tomographic scans taken of the same region in different patients, as shown in Figure 1.5, may be expected to share statistical characteristics.

By making the assumption that an image f was generated by a random process it becomes reasonable to ask what the probability of that image is. This cannot be answered unless the probability distribution characterizing the random process is known. Markov random fields provide a parametric approach to model these probability distributions.

Having a statistical model of an image allows better inferences to be made about the image and the underlying scene. These inferences may involve image analysis or they may involve inferences about restoration of the true scene. Bayes' theorem tells us how to make

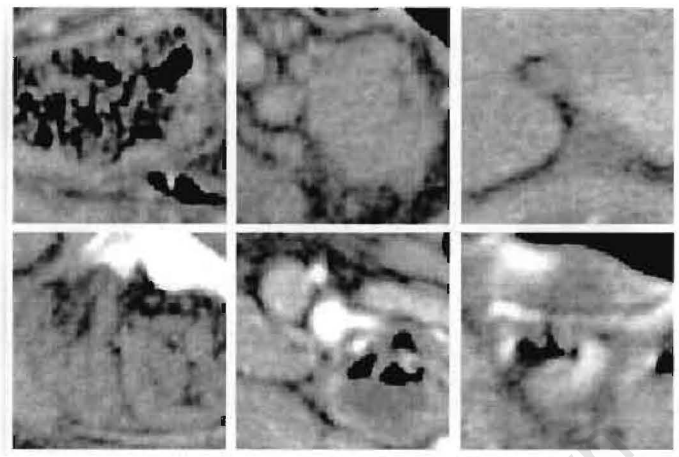


Figure 1.5: Details from tomographic scans of the torso region

these inferences as is discussed in Chapter 4.

1.7 Optimality

Using Markov random fields, many problems in image processing can be viewed as optimization problems where the aim is to find the estimate that minimizes some cost function.

For Bayesian maximum *a posteriori* estimation, as discussed in Chapter 4, the aim is to find the maximum of the *a posteriori* distribution. The *a posteriori* distribution combines the likelihood distribution and the *a priori* distribution. The likelihood distribution relates the measured data to the solution space. The *a priori* distribution contains prior information about possible solutions. This distribution will be modelled as a Gibbs distribution or Markov random field.

Rather than solve the problem in this form, it is often reasonable to take the negative log of the *a posteriori* distribution as the cost function to be minimized. The negative log

of a distribution is known as the energy of the distribution.

If the cost function is strictly convex there exists only one minimum to the cost function. In this case, so called greedy optimization methods can be used that decrease the value of the cost function with each iteration until a minimum is reached. If the potential functions contributing to the energy are all convex functions, and the energy of the likelihood distribution is convex, then the energy of the *a posteriori* distribution will also be a convex function. Thus using convex potential functions allows the global maximum of the *a posteriori* distribution to be found efficiently.

A function $g : \mathbf{R}^n \rightarrow \mathbf{R}$ is convex if the following inequality holds for $0 \leq \rho \leq 1$ and the domain of g is a convex set

$$g(\rho x_1 + (1-\rho)x_2) \quad \leq \quad \rho g(x_1) + (1-\rho)g(x_2) \tag{1.19}$$

for all points x_1 and x_2 in the domain of g [47]. A function is strictly convex if strict inequality holds whenever $x \neq y$. Graphically this inequality can be explained using Figure 1.6. If all chords between two points on the graph lie above the graph, then the function is convex. Figure 1.7 shows a non-convex function for which chords can be found that intersect the function.

If the cost function is not convex it may have local minima. This makes finding the minimum of the function much more difficult than if the function was convex, especially in high dimensional spaces. Methods that converge to the global minimum in the case of a convex function, may only converge to a local minima giving suboptimal results.

In practice, how the problem is modelled is often decided by the designer rather than prescribed by the physical process. The designer must decide which effects to model, for instance, Compton scatter and the effects of polychromatic X-ray sources are not taken into account in most CT reconstruction algorithms. This approach may be well justified if

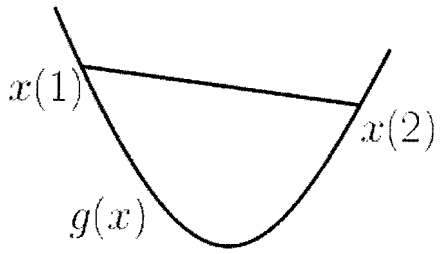


Figure 1.6: Graph of a convex function $g(x)$. The line segment between any two points on the graph stays above the graph.

these processes do not dominate the solution as they increase the computational complexity of the problem. Similarly, when it comes to modelling the *a priori* distribution of labels, convex models may be favoured due to the stability of convex models, even when non-convex models could better model the distribution. It should therefore be remembered that optimal solutions are only optimal in the sense of minimizing some cost function rather than being the best possible solution of the problem. Similarly, suboptimal solutions are local minima of the cost function. If the cost function is well chosen the minimum of the cost function should provide good quality estimates of the parameters in question.

1.8 Summary

The problem tackled in this dissertation is that of computer tomography where the measurement data is insufficient to make an estimate of sufficient quality. The approach investigated makes use of the concept of *a priori* information, that is, information known apart from measurement data. The vehicle used to capture and use this information is the Markov random field. This chapter introduced a number of concepts needed to understand Markov

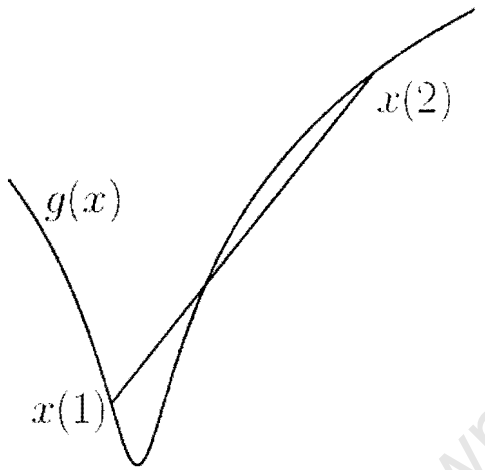


Figure 1.7: Graph of a non-convex function $g(x)$. Line segments between two points on the graph can be found such that the line segment intersects the graph.

random fields, the most important of which is the equivalence between Markov random fields and Gibbs distributions as this result is used throughout the remaining chapters.

Chapter 2

A History of Markov Random Fields

The aim of this chapter is to provide some coverage of the development and use of Markov random field theory in image processing with a focus on image restoration.

2.1 Statistical Mechanics

Much of the theory of Markov random fields was developed in the field of statistical mechanics. Statistical mechanics studies the macroscopic behaviour of bodies made up of microscopic particles such as atoms and molecules. Each particle is characterized by its state while the laws governing the interaction between particles at a microscopic level determine the macroscopic behaviour of the system.

An early example of a MRF model was the Ising model developed to study ferromagnetism in which particles can have one of two states depending on their polarization. In fact, this model has been used in image processing to model binary images [38].

Concepts such as Gibbs distributions, the temperature of a distribution, equilibrium and entropy have all found use in statistical mechanics and thermodynamics. The temperature of a Gibbs distribution plays a major role in the behaviour of the system. As the temperature

of a system is increased all configurations become equally likely and the entropy of the system is said to be high. At low temperatures the Gibbs distribution collapses, restricting the system to low energy configurations. The distribution is therefore peaked around the configurations in the state space with low energy. The effect of changing the temperature of a distribution is illustrated in Figure 2.1. The temperature of a distribution may be changed by manipulating it into the form a Gibbs distribution, as defined in Equation 1.6, from where it is a simple matter to change the temperature of the distribution.

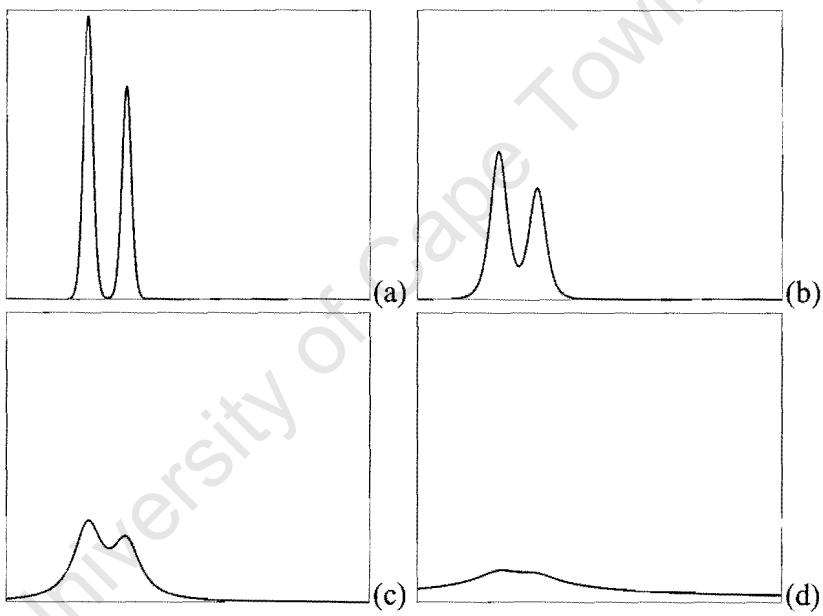


Figure 2.1: Figure shows the effect of temperature on the shape of a distribution. For image (a) $T = 20$, image (b) $T = 60$, image (c) $T = 140$ and image (d) $T = 500$

The simulated annealing optimization algorithm is another example where the inspiration of the physical world is evident. This algorithm finds low energy configurations by gradually lowering the temperature of the distribution being sampled. By starting at a high temperature and gradually lowering the temperature T, the algorithm is able to escape from

local minima [45]. The gradual lowering of temperature is designed to ensure that the system stays in equilibrium and allows very low energy configurations to be found. This is analogous to the process of annealing metal in which the metal is slowly cooled to make the metal less brittle. Slow cooling allows large crystals to form which corresponds to a low energy state.

The simulated annealing algorithm was developed by Kirkpatrick [45] and was applied to the travelling salesman problem as well as circuit layout design problems. Geman and Geman were the first to apply it to the problem of image restoration in their seminal paper *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images* [18].

2.2 Setting Out The Framework

The paper of Geman and Geman placed the use of MRFs in image processing on a firm footing by presenting a coherent way to solve image processing problems. The problem of image restoration was viewed as one of combinatorial optimization where the aim was to find the discrete labelling for a set of sites that minimized their cost function given a degraded image. Simulated annealing was presented as a method to solve the problem of image restoration in an optimal way.

Unfortunately the method of simulated annealing is computationally very expensive. This is because the temperature must be lowered very slowly at low temperatures to ensure convergence to a global minimum [18].

The paper introduced the Gibbs sampler as a means of sampling Gibbs distributions, as is required by the simulated annealing algorithm. The Gibbs sampler has a number of applications, one being that it enables image textures to be synthesized from a Markov random field model. Another important application is in Monte Carlo Markov Chain (MCMC)

methods of Bayesian estimation [7],[28].

The MRF model used in the paper of Geman and Geman was defined on a dual lattice system with an intensity process and an edge process. The edge process prevented smoothing across edge boundaries while smoothness priors were applied on the intensity process where edge processes were absent. The edge process worked by replacing the energy of the intensity process with a penalty term for sites corresponding to edges, with the magnitude of the penalty being less than the energy of the intensity process. This dual lattice model has become outdated and has been largely replaced by simpler single lattice MRFs like the Tukey potential function, see Equation 3.14.

2.3 Local Optimization

Finding globally optimal solutions remains prohibitively expensive for many image processing problems. In these cases it may be possible to obtain suboptimal estimates of sufficient quality much more quickly than the globally optimal solution. These methods search for local minima by iteratively reducing the cost of the estimate.

Besag was one of the first to present a method for finding suboptimal solutions with a method called iterated conditional modes (ICM) [6]. ICM works by updating the label at one site at time. The new label for a site is chosen so as to maximize the conditional probability for that site given the observed data and the labels at all other sites. This iterative method converges to a local minimum rather than a global minimum. Besag justified this approach by arguing that the MRF models modelled the statistical distribution at a local level and not at the global level and thus the long range statistical correlations that MRFs can introduce were in many cases undesirable. It should be pointed out that if the cost function is convex, these methods will find the globally optimal solution.

2.4 The Choice of MRF Model

The definition of Gibbs random fields allows for a wide variety of models to be generated. A number of different models have been suggested in the literature, some of which are presented in Chapter 3. Gaussian models were one of the first to be used for image processing. Gaussian models have a limited ability to model edges and this led to the adoption of discontinuity adaptive models. Many of these discontinuity adaptive models use non-convex potential functions making optimization difficult. This motivated the design of models using convex potential functions that are less difficult to solve while also producing more stable results.

The choice of MRF model also requires a neighbourhood system to be defined. In the past the neighbourhood system has often been limited to a 4 or 8 neighbourhood model for computational reasons. More recently models using much larger footprints have been developed using pyramid and wavelet decompositions [49]. The Frame model is one such example of this approach [48].

2.5 Parameter Estimation

Most Markov random field models have some parameters that change the distribution of the model. Parameter estimation is the task of selecting the parameters in a model to fit the data.

It is desirable that the *a priori* model accurately model the statistical distribution of the intensity levels when making inferences. Little has been done to address the problem of parameter estimation. It seems that in the case of image restoration, parameters are often chosen by the user. It may be argued that the MRF model is not a realistic model of the

random process generating the image, making accurate estimation of the MRF parameters a moot point. If this view is taken, the model may be seen merely as a means of adding regularization in image restoration problems, rather than a means of characterizing a random process that generated the image. Another reason why the user may estimate the MRF model parameters may be the unavailability of sample images on which to train the MRF models.

An interesting feature of the statistical framework developed in Chapter 4 for image restoration is that the *a priori* model does not change with the type and degree of degradation to the image. This is because the *a priori* model stores prior information about possible solutions which is completely independent of the measurement process. This feature is very convenient as the same model can be used in different restoration problems. It is however alarming that these models can be applied blindly to image restoration problems without taking into account how much the result is determined by the data and how much it is determined by the *a priori* model. This can be particulary serious when the model is not accurate as artifacts may be introduced by the prior model. For example, it may be reasonable to use an *a priori* model in CT reconstruction to reduce noise in the estimate. However using the same model in the case of limited angle tomography which is highly ill-posed may lead to incorrect estimates.

The obvious approach to parameter estimation if sample images are available is to choose the model parameters to maximize the likelihood of the sample data as is shown below

$$\theta^* = \arg\max_{\theta} P(f|\theta) \tag{2.1}$$

or

$$\theta^* = \arg\max_{\theta} \frac{1}{Z} \times e^{-\frac{1}{T}U(f)} \tag{2.2}$$

where the parameter vector θ is chosen so as to maximize the probability of the sample image f. However due to the high dimensionality of the configuration space, the normalizing term Z called the partition function cannot be calculated directly.

An alternative approach is to maximize a function called the pseudo likelihood (PL) as defined by Besag [6]. This approach calculates the conditional probability of each label given its neighbours and calculates the PL as the product of these conditional probabilities. This is the most widely adopted method as it has been shown to give good consistency and convergence properties as the number of sites increases [32]. It is very efficient when the dimension of labels M is low.

Images found routinely in medical applications often have 2^{12} intensity levels with 512×512 sites. The PL method becomes more computationally expensive for images with a large number of intensity levels as the conditional density at each site needs to be normalized over the M possible labels. Another complication is that for images of this type with large configuration spaces the probability of a single configuration is very small leading to potential problems associated with machine accuracy.

Parameter estimation remains one of the most difficult obstacles to using MRFs. This is especially difficult if sample images are degraded or there are no sample images and the parameters need to be estimated directly from the observed data. Parameter estimation will be discussed further in Chapter 5.

2.6 Medical Imaging

In medical imaging the adoption of MRF models has been slow due to the heavy computational requirements of MRFs. In a medical environment it is usually possible to collect enough data measurements to ensure a well posed problem. However, there are application

where MRFs have produced results of much better quality than classical methods. These include tomography applications like PET and SPECT where low photon counts make the use of accurate statistical models and *a priori* information desirable [19][14]. MRFs have also been used in the so called limited angle tomography problem in which data is not available over the full angular range of 180 degrees [2].

While objection can be made to the use of MRFs for image restoration in critical environments because incorrect *a priori* information could potentially produce image artifacts, there seems to be no such obstacles to the use of MRFs for data analysis or Computer Aided Diagnosis (CAD) where the use of prior information is unavoidable. Here MRFs could conceivably be applied to problems like the segmentation of CT data and the detection of tumours. CAD is becoming more important with the large amounts of data produced by modern diagnostic equipment as it is seldom possible for a radiologist or doctor to view all information at once making it possible to miss diagnostic information.

2.7 Present and Future Development

Future work looks set to follow the same pattern of developing more specialized models for modelling images in specific applications. As computer systems continue to get faster the adoption of MRFs should continue apace.

In addition to the large number of papers that have been published on MRFs there are a number of books on the subject although not all of them are readily available. Li's recent book provides a particularly good introduction to the varied uses of MRFs in image processing [38].

Chapter 3

Some MRF Models

In the first chapter the multi-level logistic model was presented. This model treated the labels as unordered and is thus not suitable for modelling images with a large number of intensity levels. In this chapter other MRFs will be presented that are more suitable for modelling images with a large number of intensity levels. These models treat the labels as ordered and penalize differences in the labels of neighbouring sites. These models are defined by the choice of continuous potential functions.

The input arguments for a potential function are the labels associated with the sites that fall within the clique on which the function is defined. Pairwise cliques, having two sites, are the smallest cliques to convey contextual information. Models with higher order cliques can potentially model more complex interactions between labels than models using only pairwise cliques. The models discussed in this chapter use only single and pairwise cliques.

3.1 The Auto-Normal or Gaussian Model

The auto-normal model of Besag is a type of Gaussian model [4]. The Gaussian model is defined for a continuous label set $\mathcal{L}$ by its mean and covariance terms. Its biggest advantage is that the normalizing constant can be evaluated in closed form. This contributes to the computational efficiency of this model. The covariance parameters defining the model can also be efficiently calculated [5].

The conditional probability density function for the label at a site given the labels of the neighbouring sites is given by

$$p(f_i|f_{\mathcal{N}_i}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\{f_i - \mu_i - \sum_{i' \in \mathcal{N}_i} \beta_{i,i'}(f_{i'} - \mu_{i'})\}^2}. \tag{3.1}$$

The mean or expected value of the conditional distribution for f_i, the label at site i, is given by

$$E[f_i|f_{\mathcal{N}_i}] = \mu_i - \sum_{i' \in \mathcal{N}_i} \beta_{i,i'}(f_{i'} - \mu_{i'}) \tag{3.2}$$

where $\beta_{i,i'}$ are scalar values. The variance is given by

$$\text{var}(f_i|f_{\mathcal{N}_i}) = \sigma^2. \tag{3.3}$$

When the mean value of each site is zero the conditional mean is just a weighted sum of the neighbouring pixels. The joint probability for the random field is a Gibbs distribution with the form

$$p(f) = (2\pi\sigma^2)^{-\frac{1}{2}m} |B|^{\frac{1}{2}} e^{-\frac{(f-\mu)^T B(f-\mu)}{2\sigma^2}} \tag{3.4}$$

where f is the labelling of the image in vector form, μ is a $m \times 1$ vector of the conditional means, and B is the $m \times m$ interaction matrix. B must be symmetric and positive definite for the model to be a valid probability density function. The single site and pairwise

clique potential functions for the Gaussian model are

$$V(f_i) = (f_i - \mu_i)^2/2\sigma^2 \tag{3.5}$$

and

$$V(f_i, f_{i'}) = \beta_{i,i'}(f_i - \mu_i)(f_{i'} - \mu_{i'})/2\sigma^2. \tag{3.6}$$

The auto-normal or Gaussian model is not investigated here for a number of reasons. The mean values, μ_i, are unknown and may be expected to change from image to image. Assuming the mean values are zero may be reasonable for some applications, but not for most image processing applications in which only positive pixel values are allowed.

What would be more convenient is a model that did not require estimates of the underlying mean values, but rather penalized differences in the value of labels at neighbouring sites, thus favouring smooth solutions.

3.2 Smoothness Priors

Smoothness priors are prior distributions that discourage large differences in the labels of neighbouring sites by assigning a low probability to these configurations. To do this, some metric is needed to measure the similarity of labels. If the labels are ordered then a difference operator can be defined as shown in Equation 3.7.

In practice, most images display some degree of smoothness. Smoothness priors characterize the smoothness or continuity of an image. Smoothness priors are usually defined using pairwise clique potential functions of the form given in Equation 3.7

$$V_2(f_i, f_{i'}) = g(f_i - f_{i'}) \qquad i' \in \mathcal{N}_i \tag{3.7}$$

where the function $g(\eta)$ is even so that

$$g(\eta) = g(-\eta) \tag{3.8}$$

and $g(\eta)$ is nondecreasing over the range $[0, +\infty)$ [38].

For images that are smooth, without sharp changes in intensity or discontinuities a quadratic based potential function is appropriate where β in Equation 3.9 is a scalar constant.

$$V(f_i, f_{i'}) = \beta_{i,i'}(f_i - f_{i'})^2 \tag{3.9}$$

The conditional probability for the label at a site i is given by

$$p(f_i|f_{\mathcal{N}_i}) = \frac{1}{Z} e^{-\sum_{i' \in \mathcal{N}_i} \beta_{i,i'}(f_i - f_{i'})^2}. \tag{3.10}$$

Many images do not fall into this category, exhibiting discontinuities and sharp edges. Quadratic based potential functions produce over smooth results in these cases as large changes in intensity are too heavily penalized. The obvious approach is to use potential functions that make allowance for discontinuities in the image. It turns out that there are a number of potential functions that do just that.

3.3 Discontinuity Adaptive Models

Discontinuity adaptive models are designed to allow edges to form while still providing smoothing away from the edges. Edges can be seen as the boundary between approximately flat regions. Sites falling on edges in the image are therefore classed as outliers and smoothing is not performed on them. Below are four examples of potential functions that can be used in discontinuity adaptive models.

$$g_1(\eta) = -\gamma_1 e^{-\frac{\eta^2}{\gamma_2}} + \gamma_1 \tag{3.11}$$

$$g_2(\eta) = -\frac{\gamma_1}{1 + \frac{\eta^2}{\gamma_2}} + \gamma_1 \tag{3.12}$$

$$g_3(\eta) = \gamma_1 \ln(1 + \frac{\eta^2}{\gamma_2}) \tag{3.13}$$

$$g_4(\eta) = \begin{cases} \gamma_1 \eta^2 & |\eta| \leq \gamma_2 \\ \gamma_1 \gamma_2^2 & |\eta| > \gamma_2 \end{cases} \tag{3.14}$$

The last potential function is known as the Tukey potential function and originates from the field of robust statistics [38].

The weakness of these potential functions is that they are not convex over their whole domain. The result of this is that it can be difficult to find globally optimal solutions to problems using them as gradient methods cannot be used to find optimal solutions. Another weakness of these models is that small changes in the data can lead to large changes in the result. This is highly undesirable where the robustness of the estimation is important. Often more efficient optimization methods can be used to find global minima if convex potential functions are used.

3.4 Convex Discontinuity Adaptive Models

Convex potential functions allow efficient nonlinear optimization methods to be used in place of methods like simulated annealing which are very computationally expensive. They also help to stabilize the solution of problems.

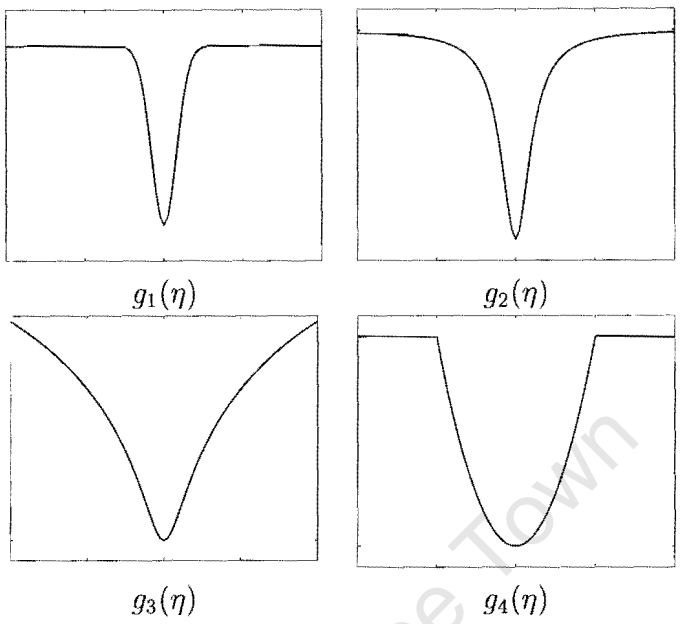


Figure 3.1: Graph of some non-convex potential functions.

Like the Tukey potential function the Huber potential function, see [11] and the references within, also originates from robust statistics and can be written as follows

$$g_5(\eta) = \begin{cases} \gamma_1 \eta^2 & |\eta| \leq \gamma_2 \\ \gamma_1(2\gamma_2|\eta| - \gamma_2^2) & |\eta| > \gamma_2 \end{cases} \tag{3.15}$$

For values of η less than γ_2 the Huber function is quadratic but for values larger that γ_2 the function is linear. The generalized Gaussian model of Bouman and Sauer [11] is given by

$$g_6(\eta) = \gamma_1 |\eta|^{\gamma_2} \tag{3.16}$$

where $1.0 \leq \gamma_2 \leq 2.0$. For $\gamma_2 = 2$ the potential function is quadratic, as γ_2 is decreased the function becomes less strongly convex. For $\gamma_2 = 1$ the function is no longer strictly convex. This can make optimization more difficult as there may be many global optima to the cost function rather than a unique optimal solution.

class	$g(\eta)$	$g'(\eta)$	$g''(\eta)$	
5	$\gamma_1\eta^2$	$2\gamma_1\eta$	$2\gamma_1$	$\lvert\eta\rvert \leq \gamma_2$
	$\gamma_1(2\gamma_2\lvert\eta\rvert - \gamma_2^2)$	$2\gamma_1\gamma_2\,\mathrm{sgn}(\eta)$	0	$\lvert\eta\rvert > \gamma_2$
6	$\gamma_1\lvert\eta\rvert^{\gamma_2}$	$\gamma_1\gamma_2\lvert\eta\rvert^{\gamma_2-1}\,\mathrm{sgn}(\eta)$	$\gamma_1\gamma_2(\gamma_2-1)\lvert\eta\rvert^{\gamma_2-2}\,\mathrm{sgn}(\eta)$	
7	$\gamma_1\ln(\cosh(\eta/\gamma_2))$	$\gamma_1\gamma_2^{-1}\tanh(\eta/\gamma_2)$	$\gamma_1\gamma_2^{-2}(1-\tanh^2(\eta/\gamma_2))$	

Table 3.1: The first and second derivatives of some convex potential functions

Another convex potential function, attributed to Green [24], is given by

$$g_7(\eta) = \gamma_1 \ln(\cosh(\eta/\gamma_2)). \tag{3.17}$$

Table 3.1 gives the first and second derivatives of the convex potential functions which are required by the maximum *a posteriori* reconstruction algorithm used in Chapter 6. Figure 3.2 gives the graph of the different convex potential functions and their derivatives.

The convex potential functions g_5, g_6 and g_7 each have two free parameters, γ_1 and γ_2. When training different models to fit sample images, as discussed in Chapter 5, setting the range of allowed values for each parameter can be difficult. This is because the magnitude of some potential functions can change by orders of magnitude with the choice of the free parameters. This can lead to problems of numerical accuracy. The approach used here is to normalize each potential function as shown by Equation 3.18. This normalization must be repeated every time γ_2 is changed and is performed after the adjustment of Equation 1.9.

$$g_N(\eta) = g(\eta) / \max_{l_i \in \mathcal{L}} g(l_i)\big|_{\gamma_1=1} \tag{3.18}$$

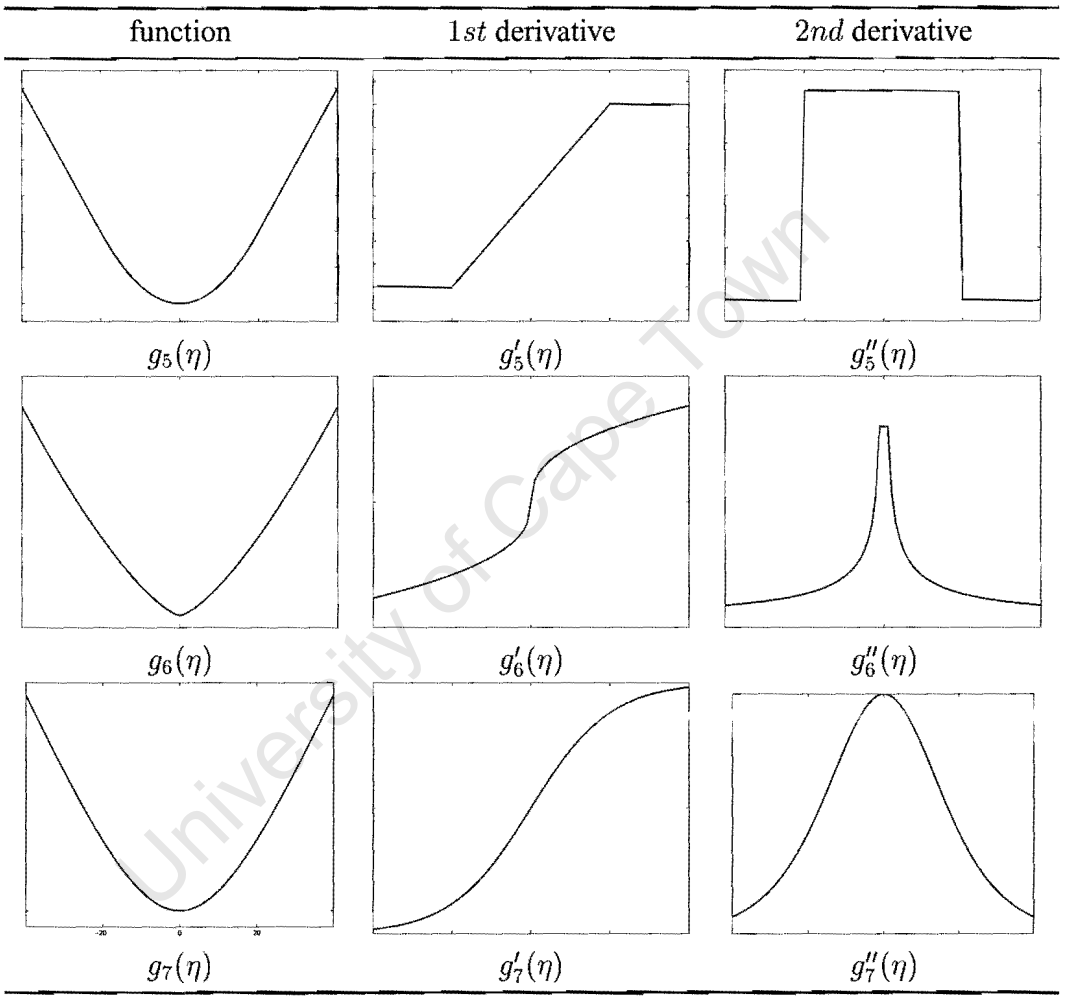


Figure 3.2: Graph of different convex potential functions and their derivatives.

3.5 Single Site Clique Potential Functions

Although the maximum *a posteriori* reconstruction algorithm in Chapter 6 does not require single site clique potential functions to be defined, it is a good idea to incorporate a single site potential function into the model when training. A single site clique potential function can be used to ensure that each configuration has a non-zero probability. Single site clique potential functions cannot convey contextual information, but only information on the relative frequency of each label.

The simplest single site clique potential function is the uniform prior. This potential function assigns the same probability to each label. It is however useful for pseudo–likelihood parameter estimation as it ensures that all configurations have a non-zero probability and because it introduces another degree of freedom into the model which allows the temperature of the distribution to be set. The uniform prior is given by

$$g_8(\eta) = \gamma_1 \tag{3.19}$$

where γ_1 is a positive constant. Equation 3.20 gives a single site clique potential function that can be used to favour labels with either small or large values.

$$g_9(\eta) = \gamma_1 \eta. \tag{3.20}$$

This potential function is more likely to find application modelling zero mean processes than for analyzing images. While this potential function introduces a bias in favour of large or small values it is a convex potential function and may therefore be used with gradient methods of image restoration.

The most general single-site clique potential assigns a weight proportional to the relative frequency of each label as was shown in section 1.5. The relative frequency of different

labels can change significantly from image to image. This makes obtaining this information *a priori* difficult. This prior may be non-convex in which case it is not suitable for use by the convex reconstruction algorithm in Chapter 6.

3.6 Summary

In this chapter a number of potential functions have been presented that can be used to form an *a priori* model. The most interesting of which for image restoration are the convex potential functions that are chosen so as not to over smooth edges. In later chapters estimating the parameters γ_1 and γ_2 will be discussed. The subscripts on the potential functions given in this chapter denote the class of the potential function and are maintained in later chapters.

Chapter 4

From Classical to Bayesian Estimation

This chapter introduces some concepts from estimation theory and looks at some of the assumptions behind using MRFs in image processing.

4.1 Introduction

In the introductory chapter, it was stated that many problems in image processing can be abstracted to one of estimating the labelling f of a set of sites denoted by $\mathcal{S}$. This chapter deals with how one makes inferences about the labelling f from the measurement data d. Two approaches are discussed, classical estimation theory and Bayesian theory. For an introduction to statistical estimation theory see [34]. For a discussion of Bayesian estimation see [31].

4.2 Classical Estimation

To make estimates of a parameter or set of parameters one needs to have a set of data measurements and an observation model with which to interpret the measurement data. The observation model can be represented in the form of a probability density distribution.

The parameters are considered to be deterministic but unknown. The measurements are corrupted by random noise. This introduces uncertainty into the observational model and allows a probabilistic approach to be taken. This model is known as the likelihood function. The likelihood function $P(d|f)$ is the likelihood of measuring the data d for the labelling f.

Once a measurement or observation model has been determined the goal is to estimate the labelling f from the probability density distribution. Just how this is done depends on the estimator used. Minimum variance unbiased estimators are generally favoured if they can be calculated. By definition minimum variance estimators have the smallest average mean square error from the true solution.

4.2.1 Maximum Likelihood Estimation

The maximum likelihood (ML) estimate is given by the mode or location of the peak of the likelihood distribution. This estimate is not optimal in the sense of being a minimum variance estimator although it is a popular choice of estimator due to the fact that the solution is always defined and the solution is often feasible to calculate. The maximum likelihood estimator gives the estimate, f^*, that maximizes the probability of the measurement data.

$$f^* = \arg\max_{f \in F} P(d|f) \tag{4.1}$$

If there is sufficient data the ML estimator gives good results. However for problems with insufficient data which are known as ill-posed, the maximum likelihood estimator has a tendency to over-fit the solution to the data leading to poor estimates. Uncertainty in the data measurements is amplified in the solution as the ML estimator has no regularization. The ML estimator makes no use of *a priori* information.

4.3 Bayesian Estimation

In classical estimation the parameters are assumed to be deterministic but unknown [34]. In Bayesian estimation the parameters may be a realization or sample from a random process that can be represented by a probability density function. In this case the *a priori* distribution may model the random process generating the samples. It is however not necessary for Bayesian estimation that the parameters being estimated be samples from a random process. In this case the *a priori* distribution represents the state of our incomplete knowledge about the parameters [31].

The *a priori* probability density function contains information about desirable solutions. This information does not depend on the observed data and is known prior to the samples being taken. Bayesian theory describes how this information can be used to obtain better solutions. It is a fundamental principle that incorporating more information into an estimator will improve the quality of the estimator. Bayes' theorem describes how to combine the likelihood function and the *a priori* probability density function in an optimal manner to form an *a posteriori* distribution containing all information about the solution.

One of difficulties of using Bayesian estimation is to obtain the prior distribution. In other words, one needs to estimate the prior probability density function before one can use it obtain the *a posteriori* distribution. MRFs may be used to model the *a priori* distribution. Once the form of the model has been selected there are usually some parameters that need to be estimated to fully define the probability density function. This problem can be approached in two ways: one can treat the parameters as missing data and use expectation maximization techniques to estimate the parameters at the same time as one estimates the solution [3][52], or one can estimate the parameters from a training set of sample images. The latter approach is discussed in Chapter 5.

The posterior probability distribution can be calculated using Bayes' theorem as follows

$$P(f|d) = \frac{P(d|f)P(f)}{p(d)} \tag{4.2}$$

where $P(d|f)$ is the conditional probability of the observations d, $p(f)$ is the *a priori* probability of the labelling f and $p(d)$ is the prior probability of making the observation d. In this work $p(d)$ will be treated as a constant.

4.3.1 Maximum A Posteriori Estimation

Once the *a posteriori* distribution has been determined, various estimates of the the labelling can be made. The maximum *a posteriori* (MAP) estimate is given by the mode of the *a posteriori* distribution.

$$f^* = \arg\max_{f \in F} P(f|d) \tag{4.3}$$

The quality of Bayesian estimates is dependant on the quality of the information stored in the *a priori* model. If the *a priori* model is valid, the MAP estimator will display better performance than the ML estimator.

Bayesian estimation allows all available information to be used in making an estimate, and thus has the potential to produce better results than classical estimation. If the problem is well posed in the sense that the data specifies a unique solution the classical and Bayesian estimates should coincide. In cases where the solution is not well posed the *a priori* distribution may add valuable information needed to make a useful estimate.

So far MRFs have been presented as a means to model the *a priori* distribution of a set of labels. In some cases MRFs may also be used to model the likelihood function. For example, a number of MRFs could be used to model different textures in an image. These

could be used to obtain a likelihood function to segment an image into regions of different texture. Another MRF could be used to model the *a priori* distribution of these regions in the image, see [50] for an example of this approach.

4.4 Summary

In this chapter, Bayes' theorem has been presented as the optimal way to update a probability density function to incorporate all available information about a set of parameters. MAP estimation was then presented as a method of inference for estimating a set of parameters from an *a posteriori* distribution.

Chapter 5

Estimating Parameters of Markov Random Fields

In this chapter a number of ways to estimate the parameters of a Markov random field are discussed. Before this can be done the form of the MRF model needs to selected. This involves selecting the neighbourhood structure and the form of the clique potential functions.

If the *a priori* model is not estimated from a set of sample images but is instead set by a user, one cannot claim to be systematically approaching the problem and the approach loses its advantage over other ad hoc methods where some *a priori* knowledge is implicit in the method.

This chapter is therefore of primary importance if a systematic approach is to be taken to image processing in general and computer tomography reconstruction in particular. Without methods of fitting MRF models to a set of data, different models cannot be compared and the question of whether it is reasonable to adopt a Bayesian approach cannot be tackled. Parameter estimation remains one of the harder problems associated with MRFs.

5.1 Introduction

When Markov random field models are used in image processing the underlying assumption is that the images of interest can be modelled by a random process. This random process is characterized by its probability distribution which in most cases will be unknown. Markov random field models provide parametric models of these probability distributions. By approximating the probability distribution using a Markov random field model, the probability distribution can be estimated from sample images. Thus the probability distribution can be estimated by estimating the free parameters of the Markov random field model.

One of the assumptions made when using MRFs is that the Markov property holds for some neighbourhood structure $\mathcal{N}$. Finding the neighbourhood structure cannot be separated from the problem of selecting clique potential functions. This is the problem of model selection. In this chapter it will be assumed that the form of the model has been previously selected. The form of the Markov model is usually chosen by the user although one form may allow for a variety of images to be modelled by changing the MRF model parameters.

The selection of the form of the MRF by the user may be motivated by a number of requirements. These may include computational requirements and modelling requirements, such as the need to model long range interaction of labels. The adoption of more sophisticated MRF models may require a greater number of sample images from which to estimate the parameters and may also require more sophisticated methods of parameter estimation [49].

5.2 Maximum Likelihood Estimate

Parameter estimation is usually based on the maximum likelihood principle where the MRF model parameters are estimated so that the distribution defined by the MRF model maximizes the probability of the sample images. Searching for the ML estimate of the MRF parameters usually requires the calculation of the likelihood of the sample images for different parameter values. Unfortunately, this is very computationally expensive. This is because calculating the likelihood of an image f given the MRF parameters θ requires the partition function Z to be evaluated. This is usually computationally unfeasible even for small images with a small number of labels or intensity levels.

One of the weaknesses of the maximum likelihood estimator is that it overfits a model to a data set if the model has sufficient modelling capacity. It is therefore important that the model have limited modelling power so that overtraining the model is not a problem. This will not be a problem for the simple models used here, although as computational power allows more complex models to be used this may become a consideration.

Given a sample image, f, the maximum likelihood estimate of the free parameters, θ^*, maximizes the conditional probability, $P(f|\theta)$, as shown in Equation 5.1.

$$\theta^* = \arg\max_{\theta} P(f|\theta) \tag{5.1}$$

This is equivalent to maximizing the log-likelihood function

$$\theta^* = \arg\max_{\theta} \ln P(f|\theta) \tag{5.2}$$

that for computational reasons may be favoured over Equation 5.1.

There are a few cases in which closed form solutions exist for the maximum likelihood estimate, however this not the case in general. The maximum likelihood parameter estimate for MRFs is generally difficult to obtain.

5.3 Pseudo-likelihood Estimate

This is probably the most common method of parameter estimation for MRF models. The method was first proposed by Besag [5]. It calculates the conditional probability for each site based on its neighbourhood. It then estimates the joint probability of a labelling as the product of these conditional probabilities as shown in Equation 5.3. The pseudo-likelihood (PL) estimate only equates to the true likelihood distribution in the trivial case in which the labels are independent.

The PL is only an approximation to the true likelihood. However, existence, uniqueness and consistency have been proved for the maximum pseudo-likelihood estimate [32].

$$PL(f) = \prod_{i \in \mathcal{S} - \partial\mathcal{S}} P(f_i | f_{\mathcal{N}_i}) = \prod_{i \in \mathcal{S} - \partial\mathcal{S}} \frac{e^{-U(f_i, f_{\mathcal{N}_i})}}{\sum_{f_i \in \mathcal{L}} e^{-U(f_i, f_{\mathcal{N}_i})}} \tag{5.3}$$

The Pseudo likelihood estimate is then given by

$$\theta^*_{PL} = \arg\max_{\theta} PL(f|\theta). \tag{5.4}$$

5.4 The Coding Method

The reason why the PL does not equate with the true likelihood is that the conditional probability of the label at each site are not independent. The coding method, also introduced by Besag [4], sidesteps the problem of dependencies between these conditional probabilities by separating them into codings so that the conditional densities in a coding are independent. The joint likelihood of a coding is then taken as the product of conditional densities, as in the PL method. This method has several weaknesses, firstly, the method does not

get maximal information from the available data, secondly, it is not obvious how to combine the estimates from different codings in an optimal manner. This method has been superseded by the PL method.

Although the coding method has been superseded for parameter estimation, the concept of dividing a lattice into codings has found other applications, like in Gibbs samplers, where a coding groups the sites that can be updated synchronously in a parallel architecture.

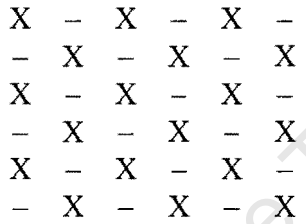


Figure 5.1: A 4 neighbour coding scheme

5.5 The Mean Field Approach

This approach to parameter estimation takes its inspiration from mean field theory in statistical mechanics. The concept behind this approach is that for a system in equilibrium the interaction between a label and its neighbours can be modelled as an interaction between the label and the mean field value.

The mean field approach takes the same form as that of the PL algorithm except that the neighbouring site labels are replaced by a mean field approximation. This decouples the conditional densities with the result that the product of the conditional densities is a better approximation to the likelihood function [16][52].

Mean field methods differ in the way the mean field values are calculated. The approach of Geiger and Girosi was to approximate the mean field by iteratively averaging neighbouring intensity levels [16]. It is not obvious how many times one should iterate this algorithm. As more iterations are used the mean field diffuses towards a uniform field. This is likely to be uninformative. There is thus a problem of scale selection when deciding on a mean field approximation.

The averaging method used by Geiger and Girosi is only valid for Gaussian MRFs where the expected value can be calculated as the weighted average of the neighbouring intensity levels.

Another approach to estimating the mean field is the saddle point approximation of Zhang [52]. This also iteratively calculates the mean field value at a site by solving the following equation.

$$\left.\frac{\partial U_i^{mf'}(f_i)}{\partial f_i}\right|_{f_i=<f_i>} = 0 \tag{5.5}$$

where

$$U_i^{mf'}(f_i) = V_c(f_i) + \sum_{j\in\mathcal{N}_i} V_c(f_i, < f_j >) \tag{5.6}$$

The solution of Equation 5.5 gives the maximum likelihood estimate of a label given its neighbours and is only an approximation to the conditional mean of the label. For Equation 5.5 to have a unique solution the potential functions must be convex. A more natural approach is to take the expected value of the conditional distribution of a site given its neighbours. For images with a large number of intensity levels the expected values could be approximated using Markov Chain Monte Carlo (MCMC) sampling methods [17][21][22].

5.6 A Cross Validation Approach

Most image processing algorithms using MRF models do not require the MRF model to be normalized. It is only when one wants to train models, using the maximum likelihood criterion, that it becomes necessary to normalize MRF models. Cross validation is appealing as a method for training Markov random fields as it does not require normalization of the MRF model.

The cross validation procedure can be summarized as follows. Estimate the value of a label using the conditional probability of the label given its neighbouring labels. Calculate the error between the label and its estimate. Repeat this procedure for all sites. Calculate the average error over the complete lattice. The average error constitutes a cost function to be minimized.

This approach makes use of the assumption that the random process generating the sample images is stationary in the sense that the properties of the random field over the lattice do not change with position. This approach may be seen as a cross validation approach, as the label of each site is left out and estimated from the other labels on the lattice [8].

The label at a site may be estimated from its conditional probability distribution in a number of ways. Using the mean of the conditional distribution as the estimator of a label is a reasonable choice as it is the estimator with the lowest variance. However this would be as computationally expensive as the PL approach. A more computationally efficient choice would be to use the mode of the conditional distribution otherwise called the ML estimate. Using this is potentially orders of magnitude faster than using the mean value estimate for images with a large number of possible labels.

5.7 Sampling Markov Random Fields

Sampling Markov random fields has a number of applications. The one of primary interest here is that of generating a set of random images from a known distribution in order to evaluate the performance of a method of parameter estimation.

A MRF sampler can be used to generate a set of images from a specific distribution or MRF model. These model parameters can then be estimated using one of the methods discussed in this chapter. The error between the true values and the estimated values can then be calculated and the bias and variance of the parameter estimator can be calculated. A complication with this approach is that the same distribution may be represented by a number of equivalent Gibbs distributions.

For this and for other applications it is imperative that the sampling method is error free. One of the major sources of error in many implementations is the use of poor pseudo random number generators [20]. Often the period of the random number generator is much too short for the large number of random numbers needed for sampling MRFs.

With the exception of Gaussian MRFs it is generally not possible to sample Markov random fields in closed form. Therefore iterative methods are used. Two samplers are discussed here, the Metropolis sampler and the Gibbs sampler. The idea behind using iterative samplers is that the iterant will converge to samples representative of the distribution.

Both the samplers discussed here sample one site at a time. The conditional density of the label at a site given the labels at the neighbouring sites is used to update the estimate of the label at the site. When all the sites have been visited one iteration of the sampler has been completed.

5.7.1 The Gibbs Sampler

The Gibbs sampler was first proposed by Geman and Geman [18]. The algorithm can be summarized as follows:

Algorithm 1 One iteration of the Gibbs sampler

repeat
Select a site i from the set $\mathcal{S}$
Sample the conditional probability density of the label at site i given the labels in the neighbourhood of site i
Replace the old label with the label just sampled
until all sites in $\mathcal{S}$ have been sampled

Implementations of the Gibbs sampler differ in the scheme used to visit each site and the manner in which the conditional probability densities are sampled.

Provided that enough iterations of the sampler are used, the order of sampling is not critical to producing valid samples [18].

Visiting sites using a raster scanning pattern may introduced artifacts into the sample images. A coding scheme may be used to prevent neighbouring sites from being sampled sequentially.

Figure 5.1 shows a coding scheme that can be used for a 4 neighbourhood model where all the sites marked with and 'x' are updated before the sites marked with a '-' are sampled. This pattern of sampling has the advantage that it is easily parallelized as all the sites with the same mark may be updated at the same time. Other sampling patterns can be used, including methods that are designed so that the transition probabilities are reversible. This consideration is important for some methods of Markov chain analysis. These methods include the random sampling of sites. For a discussion on sampling patterns see [28] or [44].

The conditional probability distribution can be calculated from the clique potential

functions as was shown by Equation 1.13. The conditional probability distribution is a univariate function and may be sampled in a number of ways. The most general approach would be to calculate the cumulative distribution and sample the inverse of it using a uniform deviate. This method has high setup and memory costs as each possible value of the label needs to be evaluated to normalize the conditional density.

A less general but often more efficient method is the rejection method [44]. The advantage of this method is that the conditional distribution needs only be known to a scale factor.

Let $f(x)$ be the probability function we want to sample and $g(x)$ be another probability density function so that $\alpha g(x) \geq f(x), \forall x \in \mathcal{L}$ where α is a scalar. The sampling procedure is given by algorithm 2.

Algorithm 2 Rejection Method

repeat
  sample X from g
  sample U from $\mathcal{U}_{[0,1]}$
until $U \leq f(X)/\alpha g(X)$
accept sample X

The probability of accepting a label in the algorithm is exactly $1/\alpha$. The closer α is to unity the more efficient the method becomes. The method does not require that α be calculated explicitly as only the ratio is important. The difficulty with this method is finding a function $g(x)$ that can be sampled efficiently. The distribution $g(x)$ may be constructed using mixtures of standard distributions like the Gaussian distribution.

5.7.2 The Metropolis Sampler

To update the current label at site i on a lattice $\mathcal{S}$ the ratio of the probability of the current label and a proposed label is calculated. The proposed label is then accepted with probability P as shown in algorithm 3. If the proposed label is rejected the current label is kept [44]. The proposed label f_i' may be taken from a uniform distribution of the possible labels although this may result in a large number of proposals being rejected.

Algorithm 3 The Metropolis sampler

generate f_i'
$\triangle U \leftarrow U(f') - U(f)$
$P \leftarrow \min\{1, e^{-\triangle U/T}\}$
if random$[0, 1) < P$ **then**
  $f_i \leftarrow f_i'$
end if

5.7.3 Comparing the Gibbs and Metropolis Samplers

It is generally not possible to say the one sampler is categorically better than the other. Much depends on the task at hand and how each algorithm has been implemented.

The Metropolis sampler is often easier to code and less computationally expensive than the Gibbs sampler. The other advantage of the Metropolis sampler is that it may be used to sample a group of sites at a time rather than a single site. This may be of benefit if there are strong interactions between labels.

The advantage of the Gibbs sampler is that it updates each site at each iteration while the Metropolis sampler may keep many of the same labels. It could thus argued that fewer iterations of the Gibbs sampler are needed to produce a sample that can be treated as independent from the initial configuration.

Chapter 6

Case Study : Transmission Tomography

In previous chapters the selection and training of MRFs has been discussed. This chapter forms a case study of how to apply a MRF model to an image processing problem. The chapter looks at how a MRF model may be used to obtain better image reconstructions in transmission tomography.

After introducing transmission tomography, the chapter presents the maximum likelihood approach that, while modelling the data measurement process statistically, does not incorporate any prior information about the reconstruction image. The maximum likelihood (ML) approach is then compared with the maximum *a posteriori* (MAP) approach that makes use of prior information.

The emphasis of this chapter will not be on the implementation of the ML and MAP reconstruction algorithms but rather on the choice of *a priori* model. Rather than selecting a model in an *ad hoc* fashion, different models are trained on sample images from a spiral CT scanner. Reconstructions from the ML and MAP algorithms are compared to determine whether the use of prior information leads to better quality reconstructions.

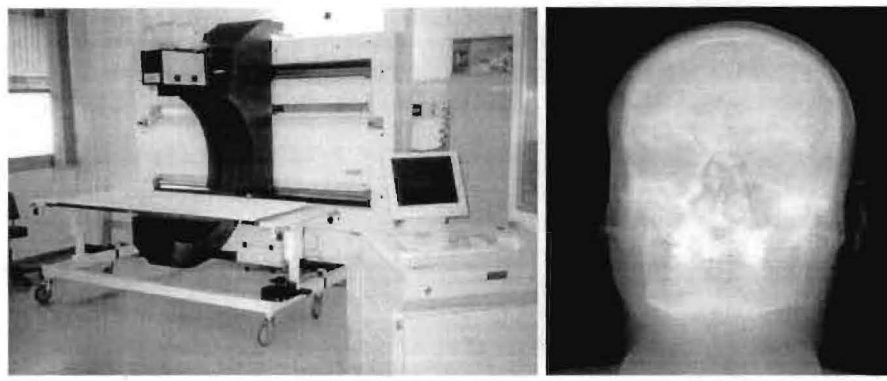


Figure 6.1: LODOX digital X-ray machine and example of an X-ray image

6.1 An Introduction to Tomography

Computer tomography allows internal anatomical detail of a patient to be examined with minimal danger to the patient. For this reason computer tomography (CT) has revolutionized medical practice since the pioneering work of Allan Cormack and Godfrey Hounsfield who together received the Nobel prize in Physiology or Medicine in 1979 [25].

Computer tomography differs from conventional X-ray scanning in that it allows cross-sectional views of a patient to be generated. This makes it possible to locate the position of anatomical structures more accurately than can be done using X-rays. It also allows small changes in density level to be seen that would be lost in X-ray images. Figure 6.1 shows a digital X-ray machine and a X-ray image. The X-ray image can be thought of as a projection of the patient's X-ray density onto an image plane. CT machines use this projection data, taking X-rays from around the patient to estimate the density at different spatial positions. Figure 6.2 shows a spiral CT machine and a CT image of the head region.

There are many varieties of tomography. They differ in modality and in application.

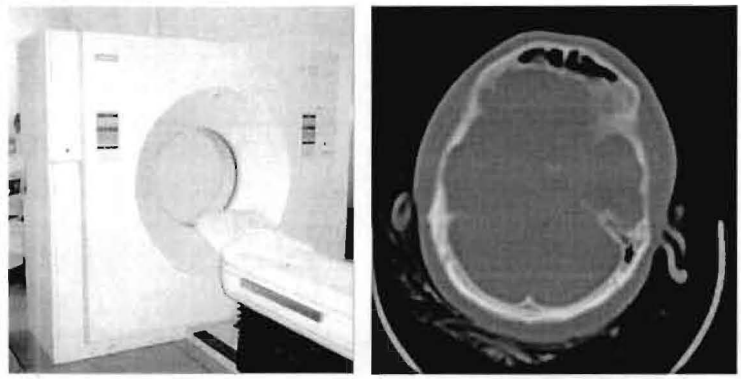


Figure 6.2: CT machine and CT image from a head study

Different modalities measure different physical attributes of the material being imaged. X-ray computer tomography images the X-ray attenuation coefficient of a material. Magnetic resonance imaging (MRI) images the resonance response of materials to a strong magnetic field [1]. Positron Emission Tomography (PET) and Single Photon Emission Tomography (SPECT) are examples of emission tomography where a radioactive isotope is administered to the patient. The isotope gives off gamma rays which are detected by a ring of detectors surrounding the patient. PET and SPECT give functional rather than structural information about the patient as the isotope concentrates in regions of high metabolic activity [1].

Tomography has also found application in other fields like nondestructive testing, radar imaging [51], seismic tomography and impedance tomography. Here only transmission tomography which includes X-ray tomography and the estimation of the X-ray attenuation map in PET are dealt with.

6.2 The Analytic Approach to Tomography

The field of computer tomography has reached a mature state of development with commercial machines able to efficiently produce good quality reconstruction images. This is largely due to efficient reconstruction algorithms based on analytic inversion formulas. These include the convolution-backprojection algorithm and the direct Fourier inversion algorithm [39]. The analytic or transform approach places emphasis on understanding the relationship between the discrete operations specified by the algorithm and the functional operations expressed by the inversion formula [37][39]. The analytical approach has produced some very efficient reconstruction algorithms that produce good results within the controlled environments in which they are used. The analytical approach has also been used to analyze sampling requirements and tackle the problem of aliasing.

At the heart of the analytical approach is the Fourier slice theorem. This theorem states that the one dimensional Fourier transform of a projection slice is equal to a slice in the two dimensional Fourier transform of the image [1].

The attenuation of a mono-energetic X-ray beam through a material with linear attenuation coefficient given by a function f is given by

$$\int f(x)dx = \int_{I_0}^{I_1} -I^{-1}dI \tag{6.1}$$

$$= \ln\left(\frac{I_0}{I_1}\right) \tag{6.2}$$

where I_0 is the number of photons emitted at the source of the X-ray beam and I_1 is the number of photons detected after passage through the material. This equation shows how the photon count data can be massaged into a form that resembles ray integrals. The nomenclature suggests that I_0 and I_1 are intensity measures rather than photon counts. The discrete nature of the photon count measurements are thus ignored.

The analytical approach does have some weaknesses: the inversion formulas assume ideal projection data without noise. Thus the analytical approach cannot lead to statistically optimal results. If the presence of noise is acknowledged it is assumed to be Gaussian in nature on the transformed data $\ln(I_0/I_1)$ and dealt with using linear filtering techniques.

The algorithms have strict sampling requirements that must be met and are thus poorly equipped to deal with changes in projection geometry. Another drawback of the analytical approach is that it does not allow prior information to be incorporated in a natural manner. Analytical methods are not used or discussed further here as they are not useful for answering the question of whether prior information is useful for transmission tomography. Instead, the series-expansion approach to tomography is taken.

6.3 The Finite Series-Expansion Approach to Tomography

Finite-series methods are based on the discrete sampling of the image domain prior to any mathematical analysis [12]. This approach allows for the data measurement and noise to be related to the image domain through a likelihood distribution. It also allows prior information to be defined on the discrete image domain and incorporated in a natural way using the maximum *a posteriori* approach.

The image domain can be modelled as a mosaic of pixels, each with constant density over their extent. It should be remembered that the pixels in CT images represent a volume in space rather than a 2D area and may more accurately be called voxels. This approach assumes that the density in a voxel is homogeneous or constant over its extent. This assumption is reasonable for most voxels although it may not be at boundaries between

different tissues. When this occurs the attenuation coefficient of the voxel is an average of the intensities of the different tissues. As the size of the voxels are increased this effect may become more apparent.

In the statistical framework for image processing that has been presented in the previous chapters a likelihood model is needed to model the relationship between the measured data and parameters to be estimated.

6.4 The Likelihood Model

The likelihood model relates the measurement data to the solution space. The more accurate the likelihood model is, the more accurately the solution can be estimated. In practice the likelihood model is limited by what is computationally feasible and by mathematical tractability. The likelihood model usually falls far short of a complete description of the measurement process.

The measurement data in CT is X-ray data. The X-ray process can be modelled at a number of different levels of complexity. The dominant effect in X-ray tomography is the absorption of X-ray photons. This occurs through what is known as the photo-electric effect. Photo-electric absorption occurs when an X-ray photon passes all its energy to an inner electron of an atom [23]. Different materials have different absorption levels. The likelihood that a photon will be absorbed by the material determines the attenuation coefficient of a material. The goal of CT is to reconstruct an image of the attenuation coefficients called an attenuation map.

In most likelihood models the X-ray beam is assumed to be mono-energetic, consisting of photons of the same energy [15][36][11][14]. This assumption allows the X-ray absorption in a region to be characterized by a single value. X-ray tubes generate poly-chromatic

X-ray beams consisting of photons with a range of energy values. The attenuation coefficient of materials changes with the energy of the X-ray photons. Modelling the effects of a poly-chromatic X-ray source in a likelihood model requires that the absorption coefficients of the different materials be known for the different photon energies and that the energy profile of the X-ray source be known [27]. This would greatly increase the complexity of the likelihood model and thus the effect of polychromatic X-ray sources is seldom modelled, even though assuming a monochromatic source may lead to beam hardening effects, including streaking and cupping in the reconstructed image [1].

In addition to the photo-electric effect there are also other interactions by which an X-ray beam is attenuated. Compton scatter is the most significant for computer tomography. Compton scattering occurs when an X-ray photon strikes an outer electron. The electron absorbs some of the photon's energy and the photon is deflected from its original path. Compton scatter is not as dependent on the energy of the X-ray photons as the photo-electric effect. Compton scatter introduces a bias into the measurement data. This affects the rays with low photon counts more than rays with large photon counts. Modelling Compton scatter accurately would be computationally expensive as the probability of photons travelling along many different paths would have to be evaluated. Because developing an accurate likelihood model is not the primary interest, Compton scatter and other absorption effects are not modelled here. The primary interest is that of developing an accurate *a priori* model with the aim of making more accurate tomography reconstructions.

The likelihood model can also be used to model noise in the detector. In a CCD this could be modelled as Gaussian thermal noise. Again, this is not modelled here as it is not of primary interest. The point of highlighting some of these phenomena is that the likelihood model is not determined solely by the objective physical nature of the physical process, but

also by subjective choices made by the user. The likelihood model, like all physical models, is merely a description of the actual physical process. The use of *a priori* information may reduce the effect of secondary effects not modelled by the likelihood model.

The absorbtion of X-rays is a probabilistic process. The probability of an X-ray photon being absorbed is related to the X-ray attenuation coefficient of the medium through which the photon is travelling. The probability of an X-ray photon reaching a detector from its source through a medium with X-ray absorbtion density μ is given by the limiting frequency

$$\frac{I_1}{I_0} = e^{-\int \mu \, ds}. \tag{6.3}$$

This is just the exponential attenuation law rewritten in a form that highlights the probabilistic nature of X-ray absorbtion.

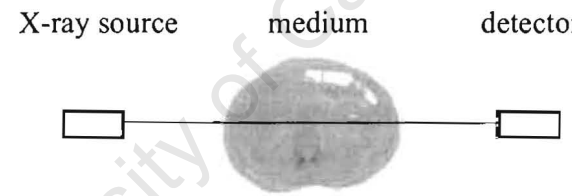


Figure 6.3: Illustration of X-ray photon travelling through medium with absorbtion density μ.

Making the assumption that the X-ray medium can be modelled by voxels with homogenous density, Equation 6.3 can be rewritten in discrete form as

$$E\left[\frac{I_1}{I_0}\right] = e^{-\sum l_{ij}\mu_j} \tag{6.4}$$

where $E\left[\frac{I_1}{I_0}\right]$ is the expected value of the ratio $\frac{I_1}{I_0}$, μ_j is the linear absorbtion coefficient at site j and l_{ij} is the projection weight for the intersection of ray i with site j. The most

obvious way to calculate the weighting coefficients l_{ij} is as the length of the ray intersecting the pixel as shown in Figure 6.4(a).

This method is appealing in that the weights have the dimension of length which agrees with Equation 6.3. The weights can also be calculated efficiently using standard line clipping algorithms [30].

In practice this model is not ideal for modelling tomographic projection for two reasons. The first is that the sampling requirements when using infinitely thin beams are very great. The second is that X-ray sources and detectors have non-negligible width.

A better approach is to model each ray as a rectangular tube [1], as shown in Figure 6.4 (b). This method leads to more stable solutions than the line intersection method although this method may not be able to model the physical geometry of the detector and source as accurately as wished. A third method and the one used here is to calculate the projection weights as the intersection of a quadrilateral with each pixel. This method allows for the geometry of the CT scanner to be accurately modelled. This method is shown in Figure 6.4(c).

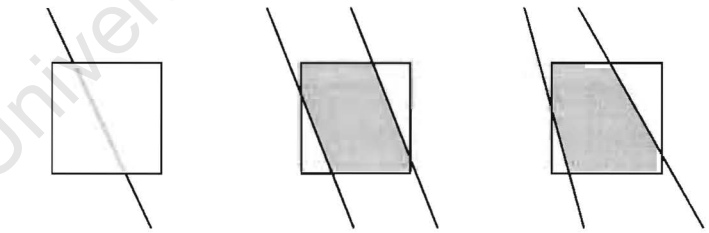


Figure 6.4: Different methods of calculating projection weights. Shaded gray regions represent the intersection of a square pixel with a projection ray. The attenuation coefficient μ is assumed to be constant over the full extent of the pixel or voxel.

6.4.1 Modelling Noise in X-ray Data

Statistical methods require the measurement noise in the data to be quantified. If the process is a discrete counting process a natural candidate is the Poisson model. If the process is continuous, a Gaussian model would probably be a more appropriate candidate. The continuous nature of the Gaussian model and the discrete nature of the Poisson model make comparison difficult. However, for high counts the shape of the Poisson distribution is very close to that of the Gaussian distribution. By sampling the Gaussian distribution, the similarity between the two probability mass functions can be calculated. The term probability mass function refers to a discrete probability distribution whereas the term probability density function is often reserved for continuous probability distributions. The Gaussian assumption does have some computational advantages and has therefore been adopted by some researchers [29]. In practice assuming Gaussian noise on the photon count data may be reasonable because of the detectors characteristics. Some detectors do not count individual X-ray photons but rather a charge proportional to the number of photons reaching the detector. Figure 6.5 shows an illustration of a digital X-ray detector. X-ray photons strike the scintillator causing a cascade of light photons to be emitted that are detected by the CCD.

While Gaussian models may be adopted in the quest for faster reconstructions algorithms, this was not the primary concern here. The more statistically correct Poisson model was therefore used.

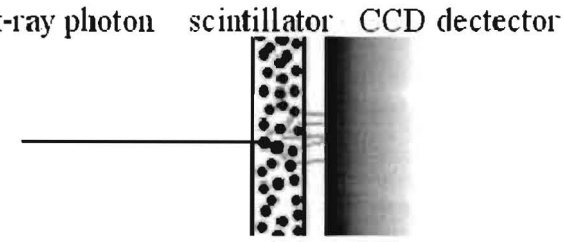


Figure 6.5: Illustration of a digital X-ray detector. X-ray photons strike the scintillator causing a cascade of light photons to be emitted that are detected by the CCD.

6.4.2 Some Properties of the Poisson Distribution

The Poisson distribution is a discrete distribution useful for modelling noise in some imaging applications. It has two distinctive properties, the first is that the distribution is restricted to positively valued integers. This is useful for modelling counting processes like photon counts in charge coupled devices (CCDs) where negative values are not feasible. The second property is that the variance changes with the mean. This is in contrast to the assumption of a Gaussian noise model with variance that does not change with the mean of the variate. The big difference between Poisson noise and other types of noise is that Poisson noise is dependent on the data whereas with other distributions, noise may be treated as an independent additive or multiplicative component.

The distribution of a non-negative, integer valued random variable Z following a Poisson distribution with mean and variance λ is given by

$$P(Z = k) = e^{-\lambda}\frac{\lambda^k}{k!}. \tag{6.5}$$

The Poisson distribution for different values of λ are shown in Figure 6.6. For low values of λ the distribution is skewed so that it is not symmetric around its mean. At higher values of λ the Poisson distribution can be approximated by quantizing a Gaussian

distribution. In fact this approximation is used by some routines for generating Poisson numbers.

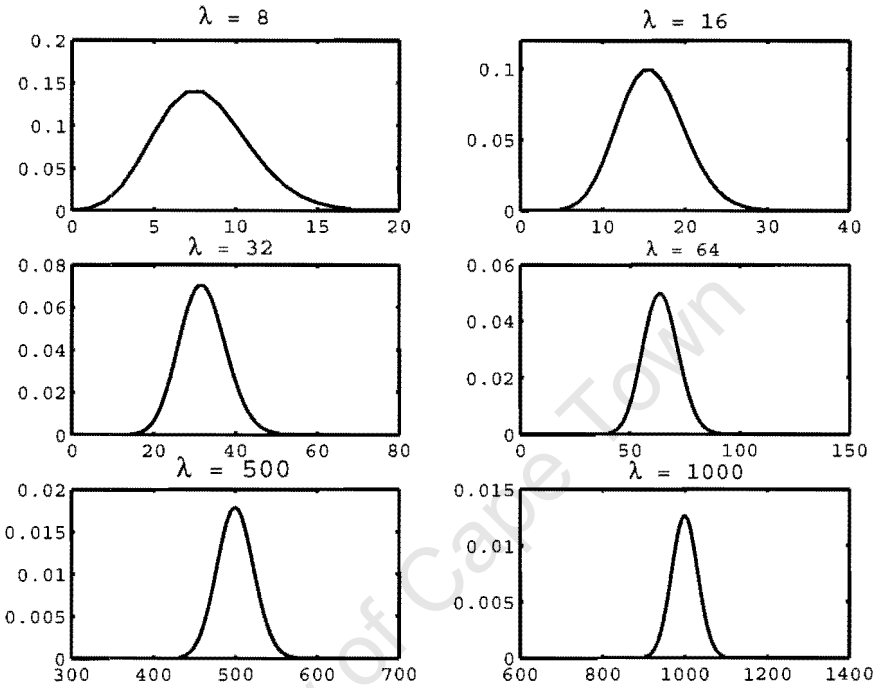


Figure 6.6: The Poisson distribution, $P(Z = k)$, where k is a positive integer, for different values of λ. To show the shape of the distribution clearly the distributions have been drawn as continuous functions, however it should be remembered that the Poisson distribution is discrete, being restricted to positive integer values.

6.4.3 Deriving the Likelihood Model

Let d_i be the expected number of photons leaving the source along ray i. The expected number of photons to reach the detector is then $d_i e^{-\sum l_{ij}\mu_j}$. Assuming that the X-ray source

is Poisson in nature the likelihood function is given by

$$G(Y,\mu) = \prod_{i=1}^{M} e^{-d_i e^{-\sum l_{ij}\mu_j}} \frac{\left(d_i e^{-\sum l_{ij}\mu_j}\right)^{Y_i}}{Y_i!}. \tag{6.6}$$

The maximum likelihood estimate of absorption coefficients μ maximizes the likelihood given photon count data Y. This is equivalent to maximizing the log-likelihood $L(\mu)$.

$$\begin{aligned} L(\mu) &= \ln\Big(G(Y,\mu)\Big) && (6.7)\\ &= \sum_i \left\{ \ln\left(e^{-d_i e^{-\sum l_{ij}\mu_j}} \frac{\left(d_i e^{-\sum l_{ij}\mu_j Y_i}\right)^{Y_i}}{Y_i!} \right) \right\} && (6.8)\\ &= \sum_i \left\{ \ln(e^{-d_i e^{-\sum l_{ij}\mu_j}}) + Y_i \ln(d_i e^{-\sum l_{ij}\mu_j}) + \ln\left(\frac{1}{Y_i!}\right) \right\} && (6.9)\\ &= \sum_i \left\{ -d_i e^{-\sum l_{ij}\mu_j} - Y_i \sum l_{ij}\mu_j + Y_i \ln(d_i) - \ln(Y_i!) \right\} && (6.10) \end{aligned}$$

The last two terms do not depend on the absorption coefficients μ and can therefore be ignored when estimating the absorbtion coefficients that maximize the likelihood. The maximum likelihood estimate μ^* is then given by

$$\mu^* = \arg\max \sum_i \left\{ -d_i e^{-\sum l_{ij}\mu_j} - Y_i \sum l_{ij}\mu_j \right\}. \tag{6.11}$$

Equation 6.11 can be solved by a number of different algorithms. The one adopted here is called the Convex algorithm of Lange [36] and will presented in section 6.11.

6.5 Outlining the Experimental Procedure

When testing reconstruction algorithms one can choose to either simulate the projection data or to use projection data from a CT machine. While the final goal must always be to perform reconstructions on real data, the testing of an algorithm may be more easily accomplished on simulated projection data.

It can be difficult to evaluate the results obtained using real data as one does not have a reference against which to measure the reconstruction quality. This difficulty may be overcome by using a phantom object with known dimensions and physical properties. This solution is not ideal in this case, because the phantom object may have different statistical properties to real CT images. Another method of overcoming the difficulty of a reference against which to measure the reconstruction quality is to simulate the X-ray process. This is done by sampling the likelihood model given a phantom image. This work differs from previous research in that it uses CT scans reconstructed by a spiral CT scanner as phantom images rather than simple artificial images made up of only a few intensity levels [15][11][9][36]. This was done to facilitate the development of more realistic and accurate MRF models.

This approach removes sources of error that occur when using real data. The geometry of the machine is known exactly. The efficiency of the detectors is known. Secondary attenuation effects like Compton scattering can be ignored. Other sources of error like beam hardening and patient movement are also avoided. Taking a simulation approach also allows for different geometries to be tried and tested.

The experimental procedure used to test whether the use of prior information leads to better quality reconstructions can be summarized as follows.

1. Collect a set of images that represent clean realizations of the *a priori* probability distribution. CT scans from a spiral CT scanner were used in this case.

2. Select the form of one or more MRF model which will be used to model the *a priori* distribution.

3. Estimate the free parameters of each MRF model using the sample images and select

the MRF model that best models the sample images. The pseudo-likelihood was used as a measure of fitness when estimating the free parameters.

4. Generate the projection data by sampling the likelihood model. Images from the set of CT scans were used as phantom images to define the linear attenuation coefficients μ_j in the likelihood model.

5. Estimate the original sample images from the experimental data samples using maximum likelihood estimation which does not use *a priori* information and using maximum *a posteriori* estimation which does make use of *a priori* information in the MRF model.

6. Compare the results of the two estimation procedures using the original sample images.

6.6 Defining the Projection Geometry

Whether one uses real or simulated data, defining the projection geometry is an essential step towards solving tomography problems. The variables defining the geometry will be explained here and the values used in the experiments will also be given.

There are two main architectures for the projection geometry. The first is a parallel beam geometry in which the coefficients are projected perpendicularly onto a detector. A parallel beam geometry models the configuration of a single point source and a single detector that is scanned linearly for each projection ray. This method of data collection is very inefficient and slow and is not used in clinical machines [1]. Some reconstruction methods demand parallel projection data. For these algorithms it is necessary to use resorting and

interpolation of fan beam data to put it into a form that can be used by algorithms designed for parallel beam data.

A fan beam geometry is adopted in the experiments that can easily be adjusted to model different CT machines. A fan beam geometry allows a complete projection slice to be measured at once using an array of detectors and a single X-ray source. More recently cone beam geometries have been developed that measure a number of projection slices simultaneously.

The position and number of projection slices determine the CT geometry. For the experiments in limited angle tomography an angular range of 100 degrees was used with either 10 or 20 projection slices. For the experiments in sparse angle tomography, an angular range of 180 degrees was used also with 10 or 20 projection slices. The projection slices are equally separated within the available angular range.

The reconstruction region must fall within the X-ray beam for all projection angles. A circular reconstruction region offers the largest possible reconstruction area for a given CT geometry although a square reconstruction region has been used here to match the sample images.

6.7 The Limited Angle Tomography Problem

The goal of limited angle tomography (LAT) as with all computerized tomography is to reconstruct an image of the internal structure of an object from projection data of the object.

The need to reconstruct images where the data is limited in its angular range occurs in many applications of computed tomography. Data acquisition may be limited by obstructions as in some non-destructive testing situations or by time constraints as in cardiac imaging.

There are well established algorithms for solving computer tomography reconstructions when sufficient data is available [1], however these fail in the case of limited angle tomography where there is insufficient data. The LAT problem is highly ill posed [39], and thus requires the use of *a priori* information to find reasonable solutions.

6.8 The Sparse Angle Tomography Problem

Sparse angle tomography (SAT) occurs when the number of projection slices are too few to uniquely determine the solution or prevent aliasing effects. The number of projection slices needed to prevent aliasing is dependent on the resolution at which one wants to reconstruct an image and the quality of the data. Commercial machines tend to use a large number of projection slices with 512 slices a reasonable number to perform a 512x512 reconstruction.

For the 128 by 128 pixel images used in the experiments, anything less than 100 projection slices may be considered as sparse angle tomography. The main reason why sparse angle tomography is worth pursuing is that, because less data is needed for sparse angle tomography, the radiation dose to the patient can be lowered. X-ray radiation can damage human tissue as it is ionizing radiation. Reducing the number of projection slices needed therefore reduces the X-ray dose to the patient. This consideration is especially important for patients like cancer sufferers who require regular scanning to determine the progress of their disease. In trauma situations this is less of a concern as multiple exposures are unlikely and the potential damage caused by the X-ray dose is far outweighed by the benefits of the CT scan.

6.9 Generating the Projection Data

The generation of projection data was setup as a number of experiments with different phantom images, projection geometries and photon counts. The experiments can be broken into two groups, those for LAT which are defined in Table 6.1 and those for SAT defined in Table 6.2. ML and MAP reconstruction algorithms are applied to the same experimental data in subsequent sections.

Experiment	Angular range	Number of projection slices	Series	Photon count
01	100	10	HIS	4000
02	100	10	HIS	2000
03	100	20	HIS	4000
04	100	20	HIS	2000
05	100	10	TIS	4000
06	100	10	TIS	2000
07	100	20	TIS	4000
08	100	20	TIS	2000

Table 6.1: Generation of projection data for LAT experiments. The angular range is given in degrees while the photon count is the number of photons leaving the source along a ray. The projection slices are equally spaced over the angular range.

The probability of making a measurement, Y_i, given that the number of incident photons along ray i is d_i, is given by

$$P(Y_i = k) = \frac{e^{-d_i e^{-\sum l_{ij}\mu_j}} (d_i e^{-\sum l_{ij}\mu_j})^k}{k!}. \tag{6.12}$$

By sampling this distribution for each measurement, Y_i, a complete set of measurement data can be generated. Reconstructions from a spiral CT scanner were used as phantom images to set the linear attenuation coefficients μ_j.

Experiment	Angular range	Number of projection slices	Series	Photon count
09	180	10	HIS	4000
10	180	10	HIS	2000
11	180	20	HIS	4000
12	180	20	HIS	2000
13	180	10	TIS	4000
14	180	10	TIS	2000
15	180	20	TIS	4000
16	180	20	TIS	2000

Table 6.2: Generation of projection data for SAT experiments. The angular range is given in degrees while the photon count is the number of photons leaving the source along a ray. The projection slices are equally spaced over the angular range.

6.10 Probability Modelling Approaches in Tomography

In Chapter 4 ML and MAP estimation was discussed, although the algorithms used to calculate these estimates were not. Most research into probability modelling approaches in tomography has centered around the development of algorithms for ML and MAP estimation. Probability modelling approaches model the data measurement process through a likelihood model like the Poisson model in Equation 6.11. The algorithms proposed to solve these estimation problems are all iterative in nature and are designed to converge to favourable solutions. Not all of them are guaranteed to converge to a globally optimal solution while others may have very different convergence rates. Algorithms are often better suited to either a parallel or serial computer architecture making comparisons difficult.

The Expectation Maximization(EM) algorithm provides an approach to solving Maximum likelihood problems [35][40]. In cases where the likelihood function may be difficult to maximize the EM approach suggests hypothesizing a complete data set in which the available data is embedded. If the expectation for this complete data set can be maximized

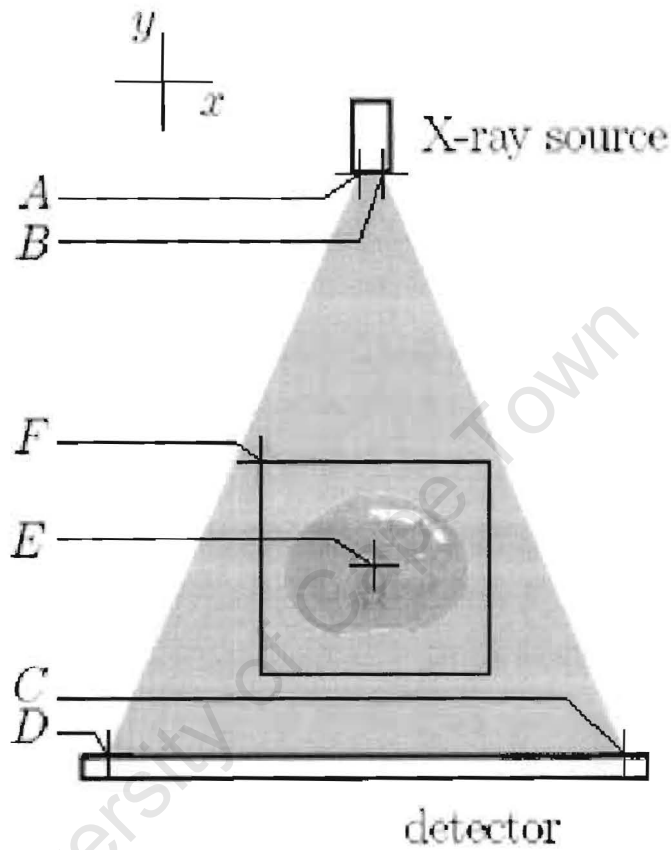
the original likelihood can also be maximized. However the algorithm does not fit the problem of transmission tomography well and more efficient algorithms have been developed [36].

A number of algorithms have been proposed that attempt to directly minimize the cost function. These include gradient methods [14], and methods like coordinate descent optimization [9].

The method chosen here to perform ML and MAP reconstructions is the Convex method of Lange [36]. This method is based on arguments proposed by De Pierro [41] for emission tomography. The method has good convergence properties and is much more efficient than the EM algorithm [36]. It is however by no means the only algorithm one can choose, there being many alternatives [15]. Some of these methods use Gaussian approximations for the noise to enable faster reconstructions [46][9][29], while others use likelihood models that allow for Compton scattering [40]. The methods discussed here only give globally optimal solutions when convex potential functions are used to model the *a priori* distribution.

There has also been some research into the development of *a priori* models suitable for tomography reconstruction [11]. These models have been designed to preserve edge information while still providing suitable regularization [2].

Because most previous work has centered around deriving estimation algorithms, the phantom images used to test them have generally been very simplistic, comprising of just a few intensity levels. While these show reconstruction errors clearly, these simple image phantoms do not display the same variety and variation as are found in real CT images. This limits their usefulness for evaluating the performance of reconstruction algorithms, especially those that use prior information, as an *a priori* model suitable for modelling simple images may not be suitable for modelling more complicated images.



label	description	value	units
A	position of X-ray source	$[0.0, 1.0]$	m
B	position of X-ray source	$[0.0, 1.0]$	m
C	position of detector	$[0.5, -0.5]$	m
D	position of detector	$[-0.5, -0.5]$	m
E	center of rotation	$[0.0, 0.0]$	m
F	position of origin of image	$[-0.2112, 0.2112]$	m

Figure 6.7: Projection geometry for specifying projection weights

6.11 The Convex Algorithm for ML Estimation

Lange and Fessler [36] discuss the Convex algorithm for transmission tomography. This method, while bearing some resemblance to the EM algorithm, does not use the concept of missing data and is less cumbersome than the EM algorithm as it does not require as many exponentiations [35]. To motivate the algorithm rewrite the log-likelihood as

$$L(\mu) = -\sum_i \psi_i(\langle l_i, \mu \rangle) \tag{6.13}$$

using the strictly convex functions $\psi_i = d_i e^{-t} + Y_i t$. As the sum of convex functions is also convex, the log-likelihood is therefore convex. In Equation 6.13 the sum $\sum_j l_{ij}\mu_j$ has been rewritten as the inner product $\langle l_i, \mu \rangle$ using matrix notation. This can be understood as integrating along a ray i over the attenuation coefficients μ_j. Terms not dependent on the attenuation coefficients have been dropped as they do not effect the optimization. Using convexity arguments the log-likelihood can be approximated by another function $Q(\mu|\mu^n)$ that relates μ to the current estimate of μ denoted by μ_n. At iteration n

$$\begin{aligned} L(\mu) &= -\sum_i \psi_i\left(\sum_j \frac{l_{ij}\mu_j^n}{\langle l_i, \mu^n \rangle}\frac{\mu_j}{\mu_j^n}\langle l_i, \mu^n \rangle\right) \\ &\geq -\sum_i \sum_j \frac{l_{ij}\mu_j^n}{\langle l_i, \mu^n \rangle}\psi_i\left(\frac{\mu_j}{\mu_j^n}\langle l_i, \mu^n \rangle\right) \\ &= Q(\mu|\mu^n) \end{aligned} \tag{6.14}$$

with strict inequality unless $(\mu_j/\mu_j^n)\langle l_i, \mu^n \rangle = (\mu_k/\mu_k^n)\langle l_i, \mu^n \rangle$ for all i and all $j \neq k$. If $\mu_j = \mu_j^n$ for all j, then 6.14 holds with equality. Inequality 6.14 is derived as a direct result of Jensen's inequality [47]. The function $Q(\mu|\mu^n)$ is designed so that the difference $L(\mu) - Q(\mu|\mu^n)$ attains a minimum of 0 at $\mu - \mu^n$. At each iteration μ^{n+1} is chosen to maximize $Q(\mu|\mu^n)$. Then the likelihood function can only increase with each iteration as

shown below

$$\begin{aligned} L(\mu^{n+1}) &= L(\mu^{n+1}|\mu^n) - Q(\mu^{n+1}|\mu^n) + Q(\mu^{n+1}|\mu^n) \\ &\geq L(\mu^n) - Q(\mu^n|\mu^n) + Q(\mu^n|\mu^n) \\ &= L(\mu^n) \end{aligned} \tag{6.15}$$

with strict inequality unless $\mu^{n+1} = \mu^n$. To maximize $Q(\mu|\mu^n)$ set

$$\begin{aligned} 0 &= \frac{\partial}{\partial \mu_j} Q(\mu|\mu^n) \\ &= -\sum_i l_{ij} \psi_i' \left(\frac{\mu_j}{\mu_j^n} \langle l_i, \mu^n \rangle \right) \\ &= -\sum_i l_{ij} [-d_i e^{-(\mu_j/\mu_j)\langle l_i, \mu^n \rangle} + Y_i]. \end{aligned} \tag{6.16}$$

Solving Equation 6.16 is guaranteed to maximize the function $Q(\mu|\mu^n)$ as the function is strictly convex. Equation 6.16 can be solved iteratively by applying Newton's method [53] for each attenuation coefficient to be estimated. Since

$$\begin{aligned} \frac{\partial^2}{\partial \mu_j^2} Q(\mu|\mu^n)|_{\mu=\mu_j} &= -\sum_i \frac{l_{ij}}{\mu_j^n} \langle l_i, \mu^n \rangle \psi_i'' \left(\frac{\mu_j}{\mu_j^n} \langle l_i, \mu^n \rangle \right) \\ &= -\sum_i \frac{l_{ij}}{\mu_j^n} \langle l_i, \mu^n \rangle d_i e^{-(\mu_j/\mu_j)\langle l_i, \mu^n \rangle} \end{aligned} \tag{6.17}$$

and

$$\frac{\partial}{\partial \mu_j} Q(\mu|\mu^n) = -\sum_i l_{ij} [-d_i e^{-(\mu_j/\mu_j)\langle l_i, \mu^n \rangle} + Y_i] \tag{6.18}$$

for $\mu_j^n > 0$, one step of Newton's method gives the approximate solution

$$\mu_j^{n+1} = \mu_j^n - \frac{\frac{\partial}{\partial \mu_j} Q(\mu|\mu^n)|_{\mu=\mu_j}}{\frac{\partial^2}{\partial \mu_j^2} Q(\mu|\mu^n)|_{\mu=\mu_j}} \tag{6.19}$$

$$= \mu_j^n + \frac{\mu_j^n \sum_i l_{ij} [d_i e^{-\langle l_i, \mu^n \rangle} - Y_i]}{\sum_i l_{ij} \langle l_i, \mu^n \rangle d_i e^{-\langle l_i, \mu^n \rangle}} \tag{6.20}$$

6.12 Models for Tomographic Data

The most general Markov models for image restoration are those that favour smooth images. This sort of model may reduce the effect of noise but will also reduce the resolution of the reconstructed image. These models do not fit computer tomography images well as they generally have sharp discontinuities in intensity at boundaries between different tissues. For instance, there is a large difference in attenuation level between regions of soft tissue and and those of bone.

One would then expect models that allow for discontinuities to perform better. These models should allow for subtle features to be reconstructed within the regions belonging to a single tissue. Some of these models are convex resulting in solutions that are solvable in a reasonable amount of time.

A more restrictive Markov random field model would be to assume that the image consists of a number of known density levels corresponding to different tissues that have been corrupted by noise. This poses tomography as a segmentation problem [13]. The Markov model would then contain information on the spatial distribution of the different levels that could be used to segment the reconstruction into a number of density levels. This type of model is non-convex making the globally optimal solution difficult to estimate.

6.13 Data Sets of Sample Images

In the chapter on parameter estimation it is assumed that sample images are available on which to train the various models. These sample images should be clean realizations of the random process. In problems like tomography one would not normally have access to sample images as these reconstructions are what one is trying to estimate. Fortunately,

commercial CT machine are able to produce good quality tomography reconstructions using conventional algorithms by gathering large amounts of projection data. The sample images used here were taken from a spiral CT scanner in digital form so that no distortion was introduced by the use film.

A series of 10 images were taken from head studies from four patients to form the Head Image Large (HIL) series. These images have dimensions 512 x 512 with 12 bits of pixel information stored in 16-bit TIFF format. A series of 10 smaller images were made from this series to form the Head Image Small (HIS) series. These images are one sixteenth the size of the HIL images and are stored in 8 bit format. This series of images are shown in appendix A.

A series of 10 images were also taken from abdominal studies from three patients to form the series Torso Image Large (TIL) and Torso Image Small (TIS).

The pixel spacing for HIL and TIL series are given in tables found in appendix A. The information is important because Markov random fields are sensitive to changes in scale. The pixel spacing in the HIS and TIS series are one quarter the length the of pixel spacing of the original images.

6.14 Defining the MRF Models

This section defines the neighbourhood structure and the clique potential functions for some proposed models. Lower and upper bounds are given for the free parameters that require estimation.

The models presented here are all based on an 8-neighbourhood model. Figure 6.8 identifies the cliques in the neighbourhood. The models are isotropic in that they do not favour a particular orientation. Because of this the potential functions for cliques $1, 2, 3, 4$

and cliques 5, 6, 7, 8 must be equal.

Model 01 is based on the Huber potential function defined in Equation 3.15. Model 02 is based on the generalized Gaussian model of Bouman and Sauer defined in Equation 3.16. The third model tested, model 03, is based on Greens potential function as defined in Equation 3.17. A uniform prior was used for clique 0 in each case, although this can be dropped for the MAP reconstruction algorithm as it has no effect on the result.

It is unusual to include parameters that change the shape of a potential function in the model as has been done here. It is far more common for the free parameters to be a set of scalar weights for a set of candidate potential functions. This can make the estimation of the free parameters more tractable. The potential function for a clique is then the weighted sum of the candidate functions. The approach taken here of including parameters that change the shape of potential functions allows the different models to be compared while also allowing for a far greater range of potential functions to be tested.

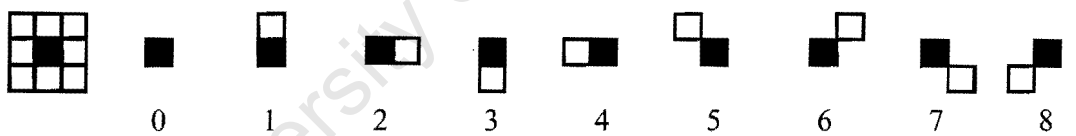


Figure 6.8: 2nd order 8 neighbourhood system and its division into cliques.

6.15 Training MRF Models on Sample Images

In this section different models are compared by training the free parameters of each model to the data sets. The PL method is used to evaluate the goodness of fit of different models on the sample images. The parameter space is limited by the definition of the potential functions and the numerical accuracy of the machine.

Clique	Potential Function	γ_1	γ_2
V_0	$g_8(\eta)$	θ_1	
V_1	$g_{5N}(\eta)$	θ_2	θ_4
V_2	$g_{5N}(\eta)$	θ_2	θ_4
V_3	$g_{5N}(\eta)$	θ_2	θ_4
V_4	$g_{5N}(\eta)$	θ_2	θ_4
V_5	$g_{5N}(\eta)$	θ_3	θ_5
V_6	$g_{5N}(\eta)$	θ_3	θ_5
V_7	$g_{5N}(\eta)$	θ_3	θ_5
V_8	$g_{5N}(\eta)$	θ_3	θ_5

Table 6.3: Definition of clique potential functions for Model 01. The parameters for each potential function are given by $\gamma 1$ and γ_2.

	θ_1	θ_2	θ_3	θ_4	θ_5
lower bound	4.0e-7	0.00	0.00	1.0	1.0
upper bound	4.0e-3	150.00	150.00	100.0	100.0

Table 6.4: Lower and upper bounds for the estimated parameters of model 01.

The results of training the convex models on the HIS and TIS image series were similar. For the parameters estimated, all three models are very similar and are all at their least convex. This is not surprising when the effect of subsampling is taken into account. The images in these two series were one sixteenth the size of the original 512x512 images. Sampling images like this, results in larger changes between neighbouring pixels. This in turn leads to less convex models, that do not penalize these changes, fitting the subsampled images better. A more surprising result was that the closest four neighbours were far more important than the diagonal neighbours which were found to contribute negligibly to the model.

The results from training the models 01 to 03 on the sample images suggest that a

Clique	Potential Function	γ_1	γ_2
V_0	$g_8(\eta)$	θ_1	
V_1	$g_{6N}(\eta)$	θ_2	θ_4
V_2	$g_{6N}(\eta)$	θ_2	θ_4
V_3	$g_{6N}(\eta)$	θ_2	θ_4
V_4	$g_{6N}(\eta)$	θ_2	θ_4
V_5	$g_{6N}(\eta)$	θ_3	θ_5
V_6	$g_{6N}(\eta)$	θ_3	θ_5
V_7	$g_{6N}(\eta)$	θ_3	θ_5
V_8	$g_{6N}(\eta)$	θ_3	θ_5

Table 6.5: Definition of clique potential functions for Model 02. The parameters for each potential function are given by $\gamma 1$ and γ_2.

	θ_1	θ_2	θ_3	θ_4	θ_5
lower bound	4.0e-7	0.00	0.00	1.0	1.0
upper bound	3.9e-3	200.00	200.00	100.0	100.0

Table 6.6: Lower and upper bounds for the estimated parameters of model 02.

non-convex model would better fit the data at the chosen resolution. All three models produce potential functions with a a similar shape for the estimated parameters, although the generalized Gaussian potential function used in model 02 proved the most likely.

The images on which the models are trained can be considered to be independent samples from the probability distribution to be estimated. To estimate the most likely parameters for a model given a series of independent sample images one must maximize the likelihood of obtaining the set of sample images for the parameters. The likelihood for a set of independently sampled images is given by

$$\prod_i P(f_{\text{image } i}|\theta) \tag{6.21}$$

Clique	Potential Function	γ_1	γ_2
V_0	$g_8(\eta)$	θ_1	
V_1	$g_{7N}(\eta)$	θ_2	θ_4
V_2	$g_{7N}(\eta)$	θ_2	θ_4
V_3	$g_{7N}(\eta)$	θ_2	θ_4
V_4	$g_{7N}(\eta)$	θ_2	θ_4
V_5	$g_{7N}(\eta)$	θ_3	θ_5
V_6	$g_{7N}(\eta)$	θ_3	θ_5
V_7	$g_{7N}(\eta)$	θ_3	θ_5
V_8	$g_{7N}(\eta)$	θ_3	θ_5

Table 6.7: Definition of clique potential functions for Model 03. The parameters for each potential function are given by $\gamma 1$ and γ_2.

	θ_1	θ_2	θ_3	θ_4	θ_5
lower bound	4.0e-7	0.00	0.00	1.0	1.0
upper bound	3.9e-3	200.00	200.00	100.0	100.0

Table 6.8: Lower and upper bounds for the estimated parameters of model 03.

with the log-likelihood given by

$$\sum_i \ln P(f_{\text{image } i}|\theta). \tag{6.22}$$

This is computationally expensive to calculate, as the PL for each image in the set needs to be evaluated in order to evaluate the joint pseudo-likelihood.

If the likelihood function and the prior function in the MAP estimation are normalized, then the estimated parameters for the *a priori* distribution can be used as is. If this is not the case a relaxation parameter λ must be introduced to determine the influence of the *a priori* distribution. This parameter is inversely proportional to the temperature of the distribution. Because the relaxation parameter λ acts as a balance between the likelihood and prior

	estimated parameters					
image	θ_1	θ_2	θ_3	θ_4	θ_5	$-\log(PL)$
HIS01	2.4528e-3	67.838	2.008	1.000	1.000	32932
HIS02	6.4003e-4	84.233	0.000	1.000	1.263	29271
HIS03	1.0000e-3	65.483	0.000	1.000	1.211	32281
HIS04	1.0000e-3	76.521	1.517	1.000	1.000	29839
HIS05	1.8965e-3	87.948	4.344	1.000	1.000	28729
HIS06	2.6669e-4	84.914	2.051	1.000	1.000	29036
HIS07	6.3934e-6	109.582	3.501	1.000	1.000	25767
HIS08	2.0835e-6	150.000	38.995	1.000	1.000	20761
HIS09	1.7264e-3	43.499	4.810	1.000	1.000	35797
HIS10	1.4378e-4	44.140	15.714	1.000	1.000	33101

Table 6.9: Results of fitting model 01 to sample images in the HIS series.

distributions it cannot be estimated directly from the sample images but must rather be evaluated in terms of the algorithm in which it is being used. In this case that algorithm is the MAP reconstruction algorithm. The relaxation parameter may be set using cross validation. Because of the expense of the MAP reconstruction algorithm cross validation is not used here. Instead the relaxation parameter was set so as to minimize the reconstruction error for the first two images in the series of ten images used in the experiments.

6.16 Hypothesis Testing

In earlier chapters the contrast between methods that make use of *a priori* information and those that don't was discussed. It was pointed out that methods that make use of *a priori* information work under that assumption that an image can be treated as a sample from a random process while non-Bayesian methods that do not use *a priori* information treat image generation as a deterministic process. The different underlying assumptions between

	estimated parameters					
image	θ_1	θ_2	θ_3	θ_4	θ_5	$-\log(PL)$
HIS01	3.9191e-3	72.332	0.000	1.000	1.895	31305
HIS02	2.4602e-3	86.745	0.000	1.000	1.999	26977
HIS03	3.9215e-3	69.634	0.000	1.000	2.000	30328
HIS04	1.0000e-3	81.625	0.000	1.000	2.000	27573
HIS05	3.2991e-3	93.579	0.000	1.000	1.997	26468
HIS06	2.0194e-3	90.103	0.000	1.000	1.556	26487
HIS07	3.9216e-3	110.421	0.000	1.000	1.184	23013
HIS08	1.0000e-3	131.768	32.170	1.000	1.000	18644
HIS09	2.3910e-3	49.728	0.000	1.000	1.880	34696
HIS10	1.0000e-3	58.752	0.000	1.000	1.513	32293

Table 6.10: Results of fitting model 02 to sample images in the HIS series.

methods that make use of *a priori* information and those that do not, make comparison of the two approaches difficult. However, it is possible to make *no* use of prior information in a Bayesian framework by adopting an uninformative or uniform prior. This distribution favours all possible solutions equally. This allows the validity of prior information to be tested in a Bayesian framework.

In order to compare the two theories a hypothesis needs to be formulated for each theory or model. The null hypothesis, H_0, is that sample images were generated by a uniform distribution. The alternative hypothesis, H_1, is that the set of sample images were sampled from a MRF, the parameters of which are given in Table 6.16.

Non-Bayesian significance tests are not suitable in this case and so a Bayesian approach using Bayes factors is taken. The Bayes factor can be calculated as follows

$$\text{Bayes factor} = B_{10} = \frac{P(H_1|Y)}{P(H_0|Y)} \tag{6.23}$$

where H_0 is the null hypothesis that the sample images in the HIS series were sampled

	estimated parameters					
image	θ_1	θ_2	θ_3	θ_4	θ_5	$-\log(PL)$
HIS01	2.4503e-4	67.677	1.887	1.000	1.000	33435
HIS02	2.4608e-3	83.966	0.000	1.000	1.000	29925
HIS03	4.2854e-7	64.969	0.000	1.000	1.000	32823
HIS04	1.7421e-4	76.258	1.336	1.000	1.000	30444
HIS05	1.0000e-3	88.210	4.129	1.000	1.000	29357
HIS06	3.9216e-3	84.738	1.885	1.000	1.000	29721
HIS07	9.9998e-4	113.364	0.000	1.000	6.637	26527
HIS08	4.0051e-7	179.886	32.411	1.000	1.000	21283
HIS09	9.9982e-4	47.119	0.000	1.000	5.063	36371
HIS10	4.4698e-7	43.903	15.407	1.000	1.000	33676

Table 6.11: Results of fitting model 03 to sample images in the HIS series.

	estimated parameters					
image	θ_1	θ_2	θ_3	θ_4	θ_5	$-\log(PL)$
HIS01	2.8770e-3	74.263	0.000	1.000	100.000	32095
HIS02	4.0000e-7	79.548	0.000	1.000	81.416	31056
HIS03	2.4373e-3	52.131	0.000	1.000	1.000	34618
HIS04	2.1613e-3	53.311	0.000	1.000	1.000	34120
HIS05	4.0005e-7	56.956	0.000	1.000	1.000	33240
HIS06	1.0000e-3	55.954	0.000	1.000	1.000	33369
HIS07	3.9191e-3	79.839	0.000	1.000	100.000	31114
HIS08	3.3077e-3	58.703	0.000	1.000	60.030	35238
HIS09	3.1771e-3	76.450	0.000	1.000	64.592	31840
HIS10	6.2252e-5	83.273	0.000	1.000	12.560	30464

Table 6.12: Results of fitting model 01 to sample images in the TIS series.

	estimated parameters					
image	θ_1	θ_2	θ_3	θ_4	θ_5	$-\log(PL)$
HIS01	4.4356e-5	71.301	0.000	1.000	1.919	31661
HIS02	3.9216e-3	75.303	0.000	1.000	1.915	30643
HIS03	2.5529e-3	53.285	0.000	1.000	2.000	33670
HIS04	1.0000e-3	54.797	0.000	1.000	1.998	33069
HIS05	2.2485e-3	58.336	0.000	1.000	2.000	32139
HIS06	2.2430e-3	57.810	0.000	1.000	1.805	32174
HIS07	3.9206e-3	75.796	0.000	1.000	1.946	30718
HIS08	1.3416e-3	58.090	0.000	1.000	1.965	34745
HIS09	3.9216e-3	72.747	0.000	1.000	1.921	31484
HIS10	1.0000e-3	78.029	0.000	1.000	1.830	30112

Table 6.13: Results of fitting model 02 to sample images in the TIS series.

	estimated parameters					
image	θ_1	θ_2	θ_3	θ_4	θ_5	$-\log(PL)$
HIS01	1.0043e-3	75.744	0.000	1.000	100.000	32325
HIS02	3.8283e-3	81.220	0.000	1.000	100.000	31310
HIS03	2.3277e-3	51.915	0.000	1.000	1.000	35000
HIS04	2.0059e-3	53.093	0.000	1.000	73.282	34513
HIS05	1.0000e-3	56.798	0.000	1.000	99.190	33645
HIS06	2.8537e-3	55.683	0.000	1.000	1.000	33786
HIS07	1.0000e-3	81.791	0.000	1.000	100.000	31325
HIS08	2.2566e-3	59.255	0.000	1.000	1.072	35460
HIS09	4.4580e-5	78.328	0.000	1.000	100.000	32039
HIS10	4.9526e-5	85.596	0.000	1.000	99.847	30687

Table 6.14: Results of fitting model 03 to sample images in the TIS series.

model	Sample Images	mean fitness	variance
model01	HIS	2.9748e+004	1.8161e+007
model02	HIS	2.7778e+004	2.1933e+007
model03	HIS	3.0364e+004	1.7926e+007
model01	TIS	3.2715e+004	2.6886e+006
model02	TIS	3.2041e+004	2.1035e+006
model03	TIS	3.3009e+004	2.8626e+006

Table 6.15: Mean and variance measures for the models 01,02 and 03.

	image	estimated parameters					
model	series	θ_1	θ_2	θ_3	θ_4	θ_5	$-\log(PL)$
model02	HIS	0.0027	76.0958	4.1904	1.0000	1.0000	282972
model02	TIS	0.0010	64.4566	0.0000	1.0000		321551

Table 6.16: LMPL estimate of the parameters of model 02 on the HIS and TIS image series.

$\log(B_{10})$	B_{10}	Evidence against H_0
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 20	Strong
> 2	> 100	Decisive

Table 6.17: Guide to interpreting Bayes factors

from a uniform distribution and H_1 is the alternative hypothesis that the sample images were generated from a MRF model.

The Bayes factor can be interpreted using the guidelines given in Table 6.17 taken from Kass and Raftery [33]. The pseudo-likelihood parameter estimation approach used evaluated the *log-likelihood* of the model in question rather than the likelihood, thus the log form of the Bayes factor will be used.

$$\begin{aligned} \log(B_{10}) &= \log\left(\frac{P(H_1|Y)}{P(H_0|Y)}\right) \\ &= \log(P(H_1|Y)) - \log(P(H_0|Y)) \end{aligned} \tag{6.24}$$

The log-likelihood for the null hypothesis can be calculated as follows

$$\begin{aligned} \log(P(H_0|Y)) &= 10\log\left(\left(\frac{1}{256}\right)^{128^2}\right) \\ &= 10 \times 128^2 \log\left(\frac{1}{256}\right) \\ &= -908521.9 \end{aligned} \tag{6.25}$$

while the log-likelihood for the hypothesis H_1 was calculated for the HIS set of sample images as -282972. Then the log of the Bayes factor, $\log B_{10} = 625549$. This result suggests that there is strong evidence to adopt the MRF model in favour of the uniform distribution for the sample images in question.

6.17 The Convex Algorithm for MAP Estimation

MAP estimation requires the maximization of the posterior function or its log. This function includes the log-likelihood function and an energy function penalizing large deviations between neighbouring pixels. The log posterior function may be written as $\Phi(\mu) = L(\mu) - U(\mu)$. The maximization function for the convex MAP algorithm is then $Q(\mu|\mu^n) - U(\mu)$ where $Q(\mu|\mu^n)$ is the maximization function derived for the convex ML algorithm in section 6.11. The function $\Phi(\mu)$ can be maximized by solving the equation.

$$0 = \frac{\partial}{\partial \mu_j} Q(\mu|\mu^n) - \frac{\partial}{\partial \mu_j} U(\mu) \tag{6.26}$$

Because the function $\Phi(\mu)$ is convex, Equation 6.26 has a unique solution. However it is difficult to evaluate in this form as we do not know μ but only its estimate μ^n. We therefore follow the approach used for the ML estimate to find a comparison function for $U(\mu)$. The energy function $U(\mu)$ is given by

$$U(\mu) = \sum_{c \in C} V_c(\mu) \tag{6.27}$$

where V_c is the clique potential on clique c. Here $U(\mu)$ can be rewritten as the sum of pairwise cliques because the uniform priors used when training the models in section 6.15 are not dependant on μ and therefore do not affect the optimization.

$$U(\mu) = \sum_{\{j,k\} \in \mathcal{N}} g_{jk}(\mu_j - \mu_k) \tag{6.28}$$

Convexity and evenness of the potential functions $g(\eta)$ together imply

$$\begin{aligned} g(\mu_j - \mu_k) &= g\Big(\frac{1}{2}[2\mu_j - \mu_j^n - \mu_k^n] - \frac{1}{2}[-2\mu_k + \mu_j^n + \mu_k^n]\Big) \\ &\leq \frac{1}{2} g(2\mu_j - \mu_j^n - \mu_k^n) + \frac{1}{2} g(2\mu_k - \mu_j^n - \mu_k^n) \end{aligned} \tag{6.29}$$

term	description
d_i	expected number of photons leaving source along ray i
Y_i	photon count for ray i
μ_j	attenuation coefficient to be estimated
l_{ij}	contribution of site j to ray i
g_{jk}	potential function for pairwise clique on sites j and k

Table 6.18: Description of terms used in sections 6.11 and 6.17.

with strict inequality unless $\mu_j + \mu_k = \mu_j^n + \mu_k^n$. Inequality 6.29 in turn yields

$$\begin{aligned} -U(\mu) &= -\sum_{\{j,k\}\in N} g_{jk}(\mu_j - \mu_k) \\ &\geq -\frac{1}{2}\sum_{\{j,k\}\in N} g_{jk}(2\mu_j - \mu_j^n - \mu_k^n) \\ &\quad -\frac{1}{2}\sum_{\{j,k\}\in N} g_{jk}(2\mu_k - \mu_j^n - \mu_k^n) \\ &= -V(\mu|\mu^n) \end{aligned} \tag{6.30}$$

The comparison function $Q(\mu|\mu^n) - U(\mu)$ is substituted by

$$\Upsilon(\mu|\mu^n) = Q(\mu|\mu^n) - V(\mu|\mu^n). \tag{6.31}$$

To find the maximum of comparison function $\Upsilon(\mu|\mu^n)$, its derivative is taken and equated to zero. Then Newton's method can be used to solve. In practice it is unnecessary to maximize $\Upsilon(\mu|\mu^n)$ at each iteration. One step of Newton's method can be used.

$$\mu_j^{n+1} = \mu_j^n - \frac{\frac{\partial}{\partial\mu_j}\Upsilon(\mu|\mu^n)|_{\mu_j=\mu_j^n}}{\frac{\partial^2}{\partial\mu_j^2}\Upsilon(\mu|\mu^n)|_{\mu_j=\mu_j^n}} \tag{6.32}$$

One iteration of the convex MAP algorithm involves solving Equation 6.32 for each site on the lattice and a number of iterations may be needed for the algorithm to converge.

The first and second derivatives of $V(\mu|\mu^n)$ are given below. It is assumed that g' and g'' are available in closed form.

$$\frac{\partial}{\partial \mu_j} V(\mu|\mu^n)|_{\mu_j=\mu_j^n} = -\sum_{\{j,k\}\in\mathcal{N}} g'_{jk}(\mu_j^n - \mu_k^n) \tag{6.33}$$

$$\frac{\partial^2}{\partial \mu_j^2} V(\mu|\mu^n)|_{\mu_j=\mu_j^n} = -2\sum_{\{j,k\}\in\mathcal{N}} g''_{jk}(\mu_j^n - \mu_k^n) \tag{6.34}$$

6.18 Estimating the Relaxation Parameter

Even when a MRF model has been trained on a set of sample images, it is often necessary to introduce a relaxation parameter λ which affects how strongly the solution is regularized by the MRF *a priori* distribution. This is necessary because the likelihood or the *a priori* distribution may not be normalized. It is not necessary that the likelihood and prior distributions are normalized so long as they are correctly balanced, hence the need for the relaxation parameter λ to balance the two distributions. The function we wish to maximize may therefore be rewritten as

$$\Phi(\mu) = L(\mu) - \lambda U(\mu). \tag{6.35}$$

The relaxation parameter λ is inversely proportional to the temperature of the *a priori* distribution. It can therefore be understood in terms of changing the temperature of the *a priori* distribution.

The relaxation parameter λ is usually set by trial and error by the user or by minimizing some measure of error. The latter approach was taken here, with the mean square error of the MAP reconstruction from its image phantom being used. Ideally λ should be set for a particular set of sample images and a particular model using a method like cross-validation

Experiment	λ
1	0.015
2	0.010
3	0.025
4	0.015
5	0.025
6	0.020
7	0.035
8	0.030

Table 6.19: Values of lambda for LAT experiments using model 02

[8]. This was found to be too computationally expensive in this case as each evaluation of the error metric requires a complete reconstruction to be made.

Instead, λ was set for each experiment by minimizing the error for just the first two images in each image series. Thus better results could be expected had the error been minimized over the whole image series or had cross-validation methods been used. Tables 6.19 and 6.20 give the λ values used for each set of reconstructions. It has been pointed out that the error metric of the mean square error tends to tends to over regularize the solution giving images an over-smoothed appearance.

6.19 Comparing ML and MAP Reconstructions

Reconstruction images for both maximum likelihood and maximum *a posteriori* reconstructions are given appendix B.

Because the projection data was calculated from a set of phantom images, the reconstruction error can easily be evaluated. The Mean Square Error (MSE) metric was used to calculate the error between the phantom images and the reconstructed images. The MSE is

Experiment	λ
9	0.010
10	0.010
11	0.020
12	0.020
13	0.030
14	0.025
15	0.035
16	0.030

Table 6.20: Values of lambda for SAT experiments using model 02

given by

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}(\mu_i^* - \mu_i)^2 \tag{6.36}$$

where m is the number of elements in the reconstructed image μ^* and μ is the original phantom image. The average MSE refers to the average MSE over a set of reconstruction images.

The mean square error for the experiments in limited angle tomography are given in Table 6.21. The MAP reconstructions for the limited angle tomography experiments show a significant reduction in MSE error over the ML reconstructions. The use of *a priori* information about the solution not only helped to reduce noise in the reconstructions but also helped to recover the underlying shape of the phantom images. This can be seen in some of the MAP image reconstructions in which the support of the reconstruction objects is more accurately recovered than in the corresponding ML reconstructions, see Figures B.11 and B.12. The support of an object in a CT scan is the area the object occupies on the image plane.

The visual quality of both the ML and MAP reconstructions was poor with neither

method able to accurately reconstruct the regions corresponding to the missing projection data. This can in large part be attributed to the very small number of projections slices used in the reconstructions.

To obtain better quality reconstructions the best course of action would be to increase the number of projection slices collected. No improvements can be made by modifying the likelihood model as it exactly matches the simulated measurement process. The *a priori* model could be further refined by adopting a first order clique potential function that more accurately models the distribution of the different attenuation coefficients. This would however make the *a priori* distribution multi-modal making estimation of the global maximum of the *a posteriori* distribution more difficult.

If taking more measurements were not feasible and refinement of the *a priori* model did not realize sufficient improvements it may be necessary to redefine the solution space. Reforming the problem as one of segmentation where only a few density levels or labels are allowed greatly reduces the dimensionality of the solution space, although the problem becomes one of combinatorial optimization rather than one convex optimization.

The visual appearance of the MAP reconstructions were susceptible to showing signs of blocking and producing false edges. This is highly undesirable in a clinical environment as false edges could lead to a misdiagnosis. Using potential functions that are more strongly convex should reduce blocking effects and the production of false edges, although this may lead to some loss of resolution in the reconstruction images.

While the reconstruction quality for both the ML and MAP reconstructions cannot be described as good, the MAP reconstructions did show a strong improvement over the ML images demonstrating that the use of *a priori* information can lead to greatly improved reconstruction images.

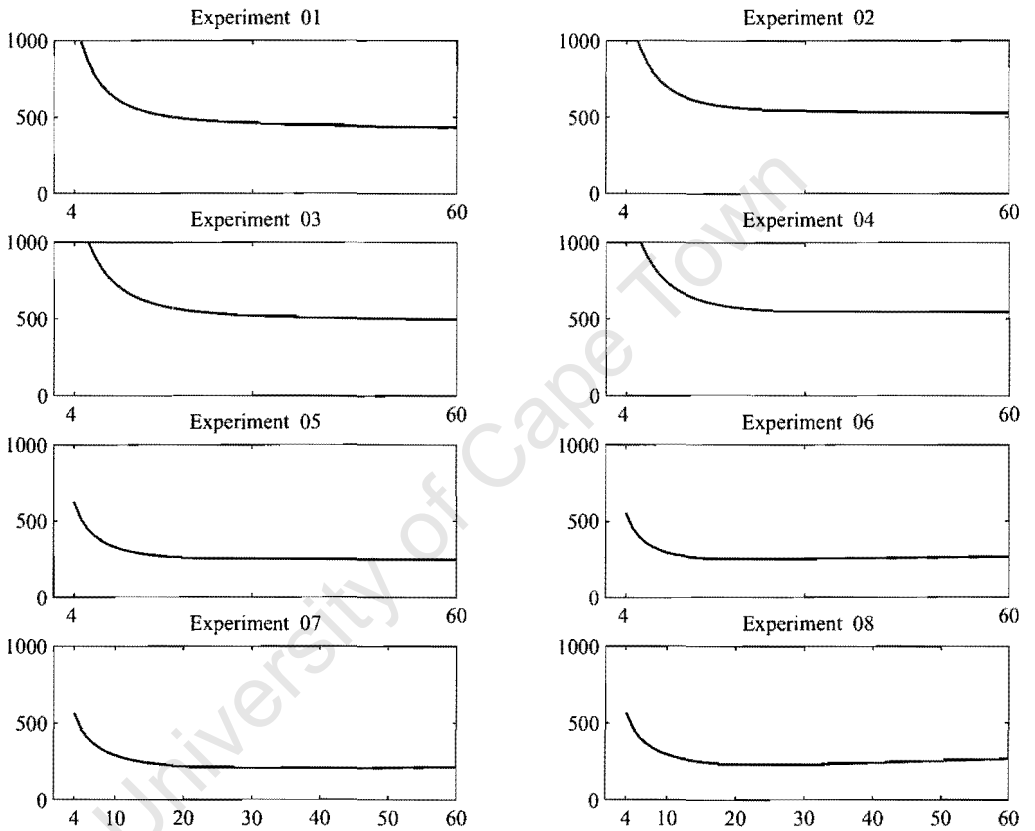


Figure 6.9: Plots showing the average MSE error against iteration number for the ML reconstructions for the different LAT experiments.

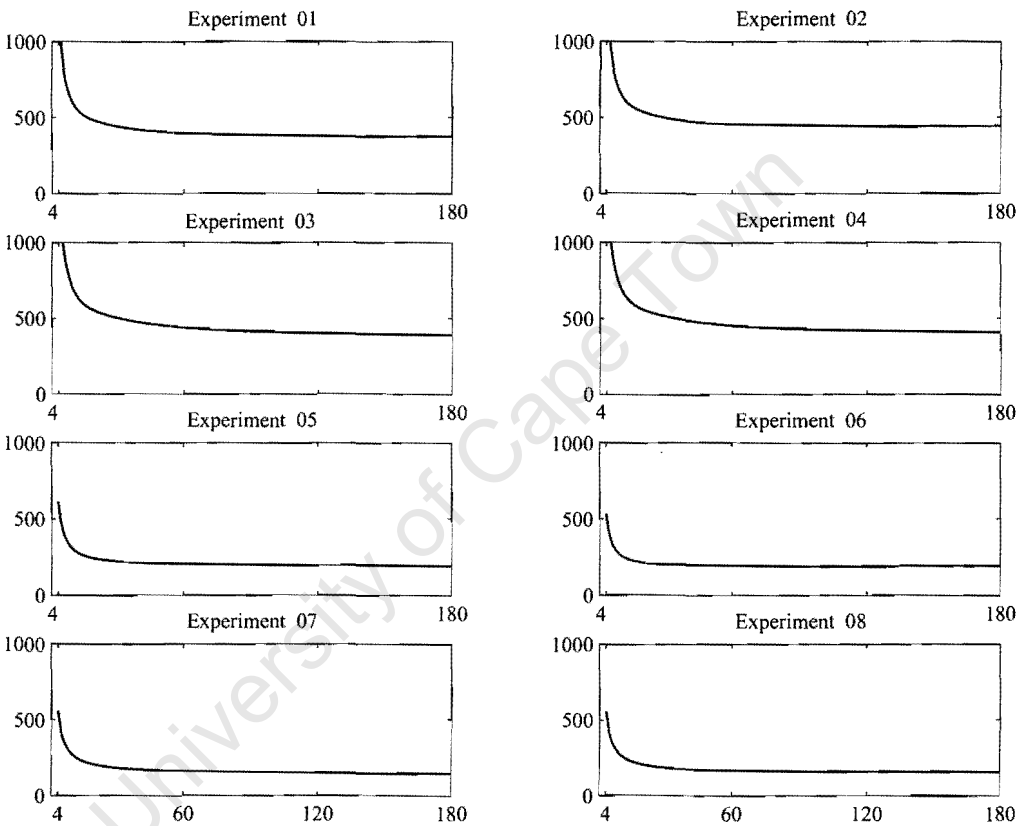


Figure 6.10: Plots showing the average MSE error against iteration number for the MAP reconstructions for the different LAT experiments.

Exp	Number of slices	Image series	Photon count	ML error mean	ML error variance	MAP error mean	MAP error variance
01	10	HIS	4000	430.88	2920.35	373.34	2177.70
02	10	HIS	2000	483.36	2096.84	406.49	1971.64
03	20	HIS	4000	443.25	2877.06	345.46	2178.31
04	20	HIS	2000	500.62	2885.61	371.63	2159.41
05	10	TIS	4000	193.55	1020.92	151.99	1210.84
06	10	TIS	2000	240.23	845.01	180.14	1266.75
07	20	TIS	4000	187.07	535.25	130.31	756.77
08	20	TIS	2000	246.53	373.34	144.58	902.91

Table 6.21: Average MSE and variance measures for the ML and MAP limited angle tomography reconstructions.

Figures 6.9 and 6.10 give the convergence results for the limited angle tomography experiments for the ML and MAP reconstruction algorithms respectively. The average mean square error between the phantom images and the reconstructions is used to measure convergence. One does not usually have a phantom image to evaluate the convergence, in these cases the change in the log-likelihood can be used as an indicator of convergence as shown in Figure 6.11.

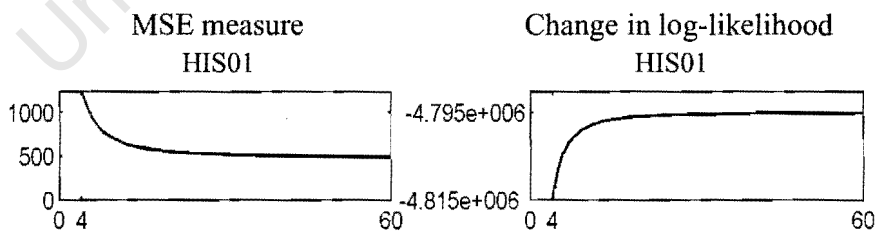


Figure 6.11: Comparison of the MSE and the change in log-likelihood as measures of convergence. The figure shows the convergence of the ML reconstruction of image 01 in experiment 01.

The ML reconstructions in the sparse angle tomography experiments showed much better results than the ML reconstructions in limited angle tomography using the same number of measurements.

In the experiments in sparse angle tomography the MAP reconstructions showed a reduction in error over the ML reconstructions in all but two of the experiments. This was due to the inaccurate selection of the λ parameter in those cases. The convergence of the sparse angle tomography reconstructions was also much faster than for the LAT reconstructions. In fact, the error of the reconstructions tended to increase or remain constant past 50 iterations of the MAP algorithm as can be seen in Figure 6.15.

The appearance of the MAP SAT reconstructions is improved over the ML reconstructions although, like the LAT experiments, use of more strongly convex potential functions may lead to visually more appealing reconstructions. The effects of using too few projection slices is evident in both the ML and MAP reconstructions. However the use of *a priori* knowledge is again justified by the results obtained as they show a significant improvement over the ML reconstructions as can be seen in Table 6.22.

Figure 6.12 allows the average results over all the LAT and SAT experiments to be compared. The SAT reconstructions far prove more accurate than the LAT reconstructions even though they represent the same X-ray dose to the patient. This shows the importance of collecting projection data over a full 180 degrees. The improvement of the MAP reconstructions can be seen for both the LAT and SAT reconstructions

Figure 6.13 shows the effect of changing the number of incident photons entering each ray. Again the MAP reconstructions prove more accurate than the ML reconstructions, but what is more surprising is that the MAP reconstructions using half the dose of the ML reconstructions still have a lower average error. What this means is that by using the MAP

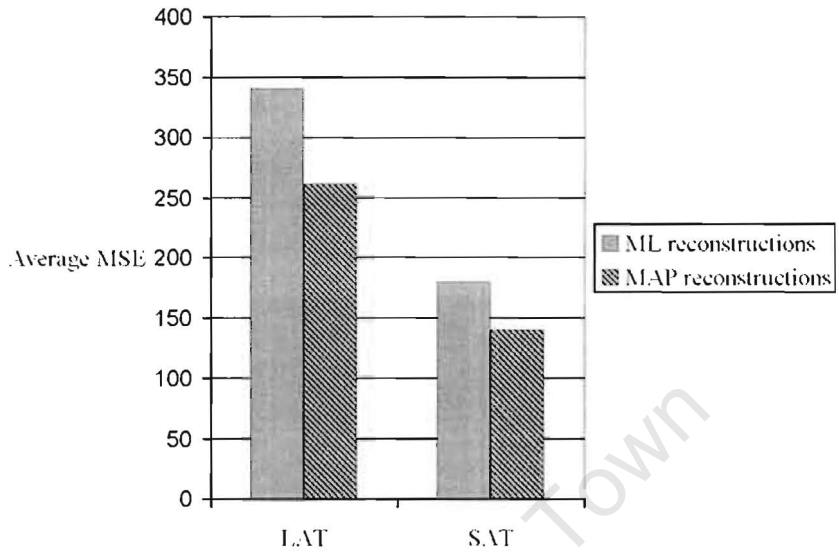


Figure 6.12: Average MSE for the LAT and SAT experiments

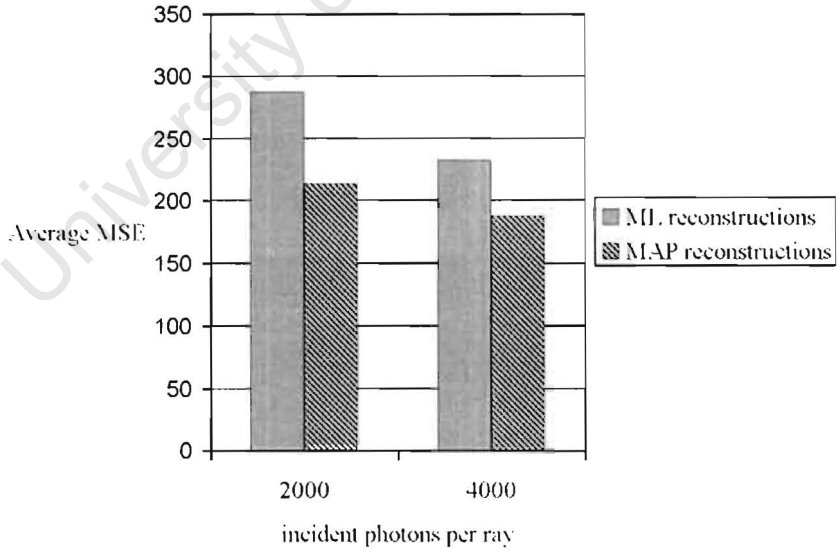


Figure 6.13: Average MSE for the different photon count experiments

Exp	Number of slices	Image series	Photon count	ML error mean	ML error variance	MAP error mean	MAP error variance
09	10	HIS	4000	184.23	2883.22	191.74	2709.66
10	10	HIS	2000	235.43	2918.92	235.46	3338.27
11	20	HIS	4000	137.74	962.32	114.60	633.33
12	20	HIS	2000	202.68	1003.46	148.78	1024.67
13	10	TIS	4000	157.09	668.56	119.68	1008.59
14	10	TIS	2000	205.96	466.10	137.80	985.80
15	20	TIS	4000	128.54	168.59	79.32	418.17
16	20	TIS	2000	191.64	245.44	95.65	398.20

Table 6.22: Average MSE and variance measures for the ML and MAP sparse angle tomography reconstructions.

algorithm the X-ray dose could be halved and still perform better than the ML algorithm.

6.20 Conclusions and Recommendations

Reconstructions using maximum likelihood and maximum *a posteriori* methods have been compared for several different problems in transmission tomography and the MAP approach using MRFs to model the *a priori* distribution were shown to give better results than the maximum likelihood method.

The theoretical framework for Bayesian image reconstruction is in a mature state. Current and future development will involve improvement to various components within this Bayesian framework. Likelihood models that more accurately model the physical image formation process are being developed to take into account effects like Compton scatter, noise in the detector and polychromatic X-ray sources. There is also work to be done in developing more sophisticated MRF models that model specific applications more accurately

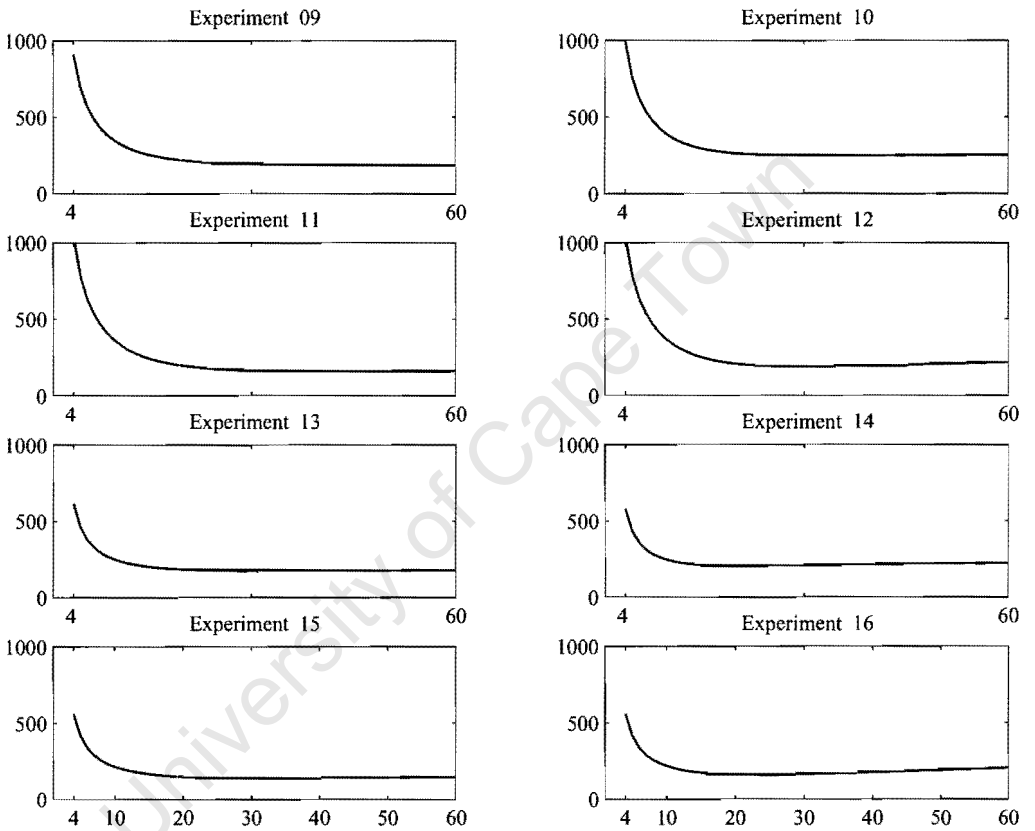


Figure 6.14: Plots showing the average MSE error against iteration number for the ML reconstructions for the different SAT experiments.

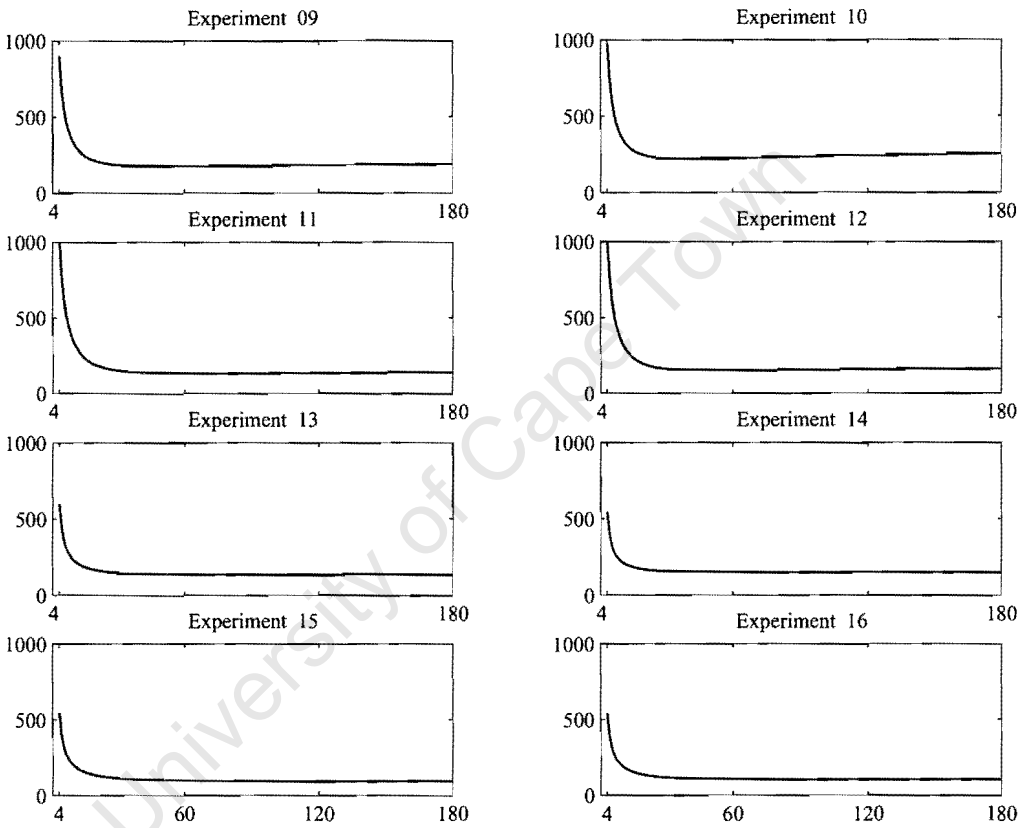


Figure 6.15: Plots showing the average MSE error against iteration number for the MAP reconstructions for the different SAT experiments.

and models that support the use of larger neighbourhoods.

The final area of development will be in the algorithms used to calculate MAP reconstructions. The biggest limitation to the adoption of these methods is the computational time involved, however improvements are being made in both computational power and the efficiency of new algorithms for MAP estimation.

Chapter 7

Conclusions and Recommendations

Markov random fields prove to be a useful tool for modelling the distribution of attenuation coefficients found in CT scans. The models were not overtrained due to their limited modelling power. Even the relatively simple models derived provided improved results in a number of different experiments over the ML method that does not take *a priori* information into account. These results are all the more impressive when one considers that the ML algorithm uses a complete statistical model of the measurement process.

The dangers of using an *a priori* model in an image restoration environment is that inaccuracy in the model may lead to artifacts in the restored or reconstructed images. For this reason convex models were investigated as convex models are more stable than non-convex models for which small changes in the input data can lead to large changes in the solution.

For the images on which the models were trained the potential functions that maximized the likelihood of the sample images were non-convex. This is not surprising as the effect of sub-sampling an image is that differences between neighbouring pixels tend to be larger. This favours non-convex models that do not penalize large changes in neighbouring pixels as harshly as convex models.

It can therefore be expected that had the images not been sub-sampled, more strongly convex potential functions would come to prominence.

For the maximum *a posteriori* reconstructions the least strongly convex convex functions were used. This was done to ensure convergence of the solution and because of the favourable properties of convex models.

The importance of training models on sample images was highlighted by the rather surprising result that the sample images were better modelled using a 4 neighbourhood model rather than an 8 neighbourhood model which one might assume to be superior due to its more symmetric structure.

One of the concerns people have with using prior models is what will happen in unusual cases. For instance, if someone with a bullet wound or extensive fractures were scanned, would the algorithm fail because the training data did not contain these examples? In special cases like these the algorithm would still work because of the very general nature of the *a priori* model that does not include specific information about attenuation coefficients or other specific information like the shape of the objects being reconstructed.

The appeal of taking a Bayesian approach has a lot to do with its modularity. One does not have to tackle the whole problem at once but rather one can approach each part as a separate problem which is then combined in the optimal manner using Bayes' theorem. For instance, one can improve an algorithm simply by adopting a more accurate likelihood or *a priori* model. One can determine whether one has made an improvement without actually having to run the whole algorithm on expensive validation testing.

Comparison of the experimental results with similar work is complicated by the different aims and goals between this and previous work. Most previous work on probabilistic approaches to transmission tomography has concentrated on the development of algorithms

for calculating the MAP estimate with the goal of designing algorithms with better convergence properties. It must be remembered that all these algorithms should give the same solution given the same likelihood and prior distributions.

The aim of this work was not to develop another MAP algorithm, but rather to develop the *a priori* model used in the MAP estimation. Where most previous work has used simple image phantoms, this work has used real CT scans as phantom images. This has allowed the development of more accurate models to model the distribution of attenuation coefficients found in real CT images. It was found that MRFs can be used to model the distribution of attenuation coefficients found in real CT scans and that these models can be used to make better MAP estimates. The results affirm the importance of previous work in developing the algorithms needed for MAP estimation in transmission tomography.

There is still much work to be done especially in the design of non-convex optimization routines. There is also work to be done in the field of training MRFs with larger regions of support. With a few caveats imposed by computational and mathematical tractability, MRFs offer a very useful tool for modelling the *a priori* distributions of images within a Bayesian framework.

Appendix A

Data Sets

This appendix describes the sample data used in Chapter 6. The sample images were taken from a spiral CT scanner. The images generated by the CT scanner had 12 bits of information and dimensions of 512 x 512 pixels. Smaller copies were make from these to reduce computational loads. These smaller images have dimensions of 128 x 128 pixels and have been quantized to 8 bit images.

A series of 10 images were taken from head studies from four patients to form the Head Image Large (HIL) series. A series of 10 smaller images were made from this series to form the Head Image Small (HIS) series. These images are one sixteenth the size of the HIL images and are stored in 8 bit format. A series of 10 images were also taken from abdominal studies from three patients to form the series Torso Image Large (TIL) and Torso Image Small (TIS).

The pixel spacing for HIL and TIL series are given in tables A.2 and A.3. This information is important because Markov random fields are sensitive to changes in scale. The pixel spacing in the HIS and TIS series are one quarter the length the of pixel spacing of the original images.

Data Set	numbering	description	image dimensions	bit depth	storage
HIL	01 - 10	head study	512 x 512	12	16-bit TIFF
HIS	01 - 10	head study	128 x 128	8	8-bit TIFF
TIL	01 - 10	torso study	512 x 512	12	16-bit TIFF
TIS	01 - 10	torso study	128 x 128	8	8-bit TIFF

Table A.1: Summary of image data sets

Image	pixel spacing (mm)	patient
HIL01	0.45117188	a
HIL02	0.45117188	a
HIL03	0.46289063	b
HIL04	0.46289063	b
HIL05	0.37695313	c
HIL06	0.37695313	c
HIL07	0.37695313	c
HIL08	0.37695313	c
HIL09	0.40039063	d
HIL10	0.40039063	d

Table A.2: Pixel spacing for images in the HIL series

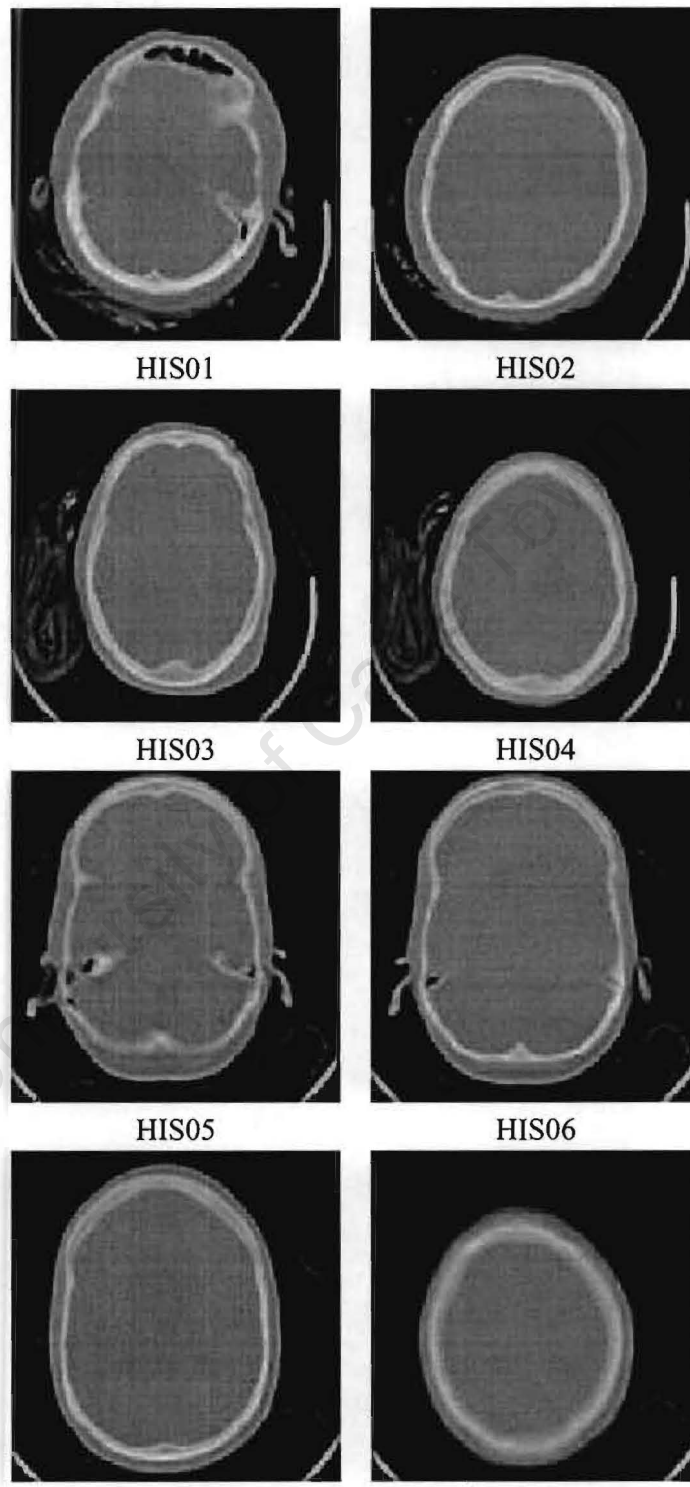
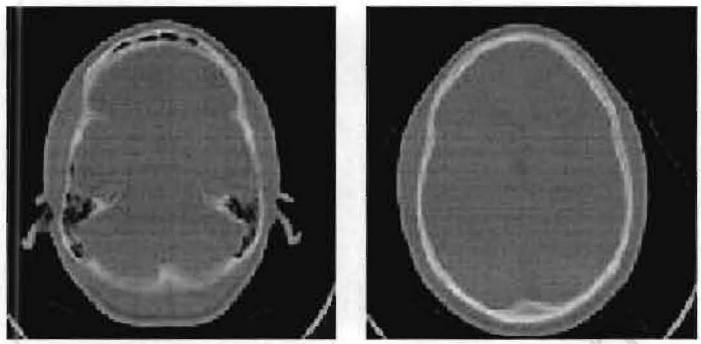


Figure A.1: Images in the HIS series



HIS09 HIS10

Figure A.2: Images in the HIS series

Image	pixel spacing (mm)	patient
TIL01	0.61914063	e
TIL02	0.61914063	e
TIL03	0.7421875	f
TIL04	0.7421875	f
TIL05	0.7421875	f
TIL06	0.7421875	f
TIL07	0.45703125	g
TIL08	0.45703125	g
TIL09	0.45703125	g
TIL10	0.45703125	g

Table A.3: Pixel spacing for images in the TIL series

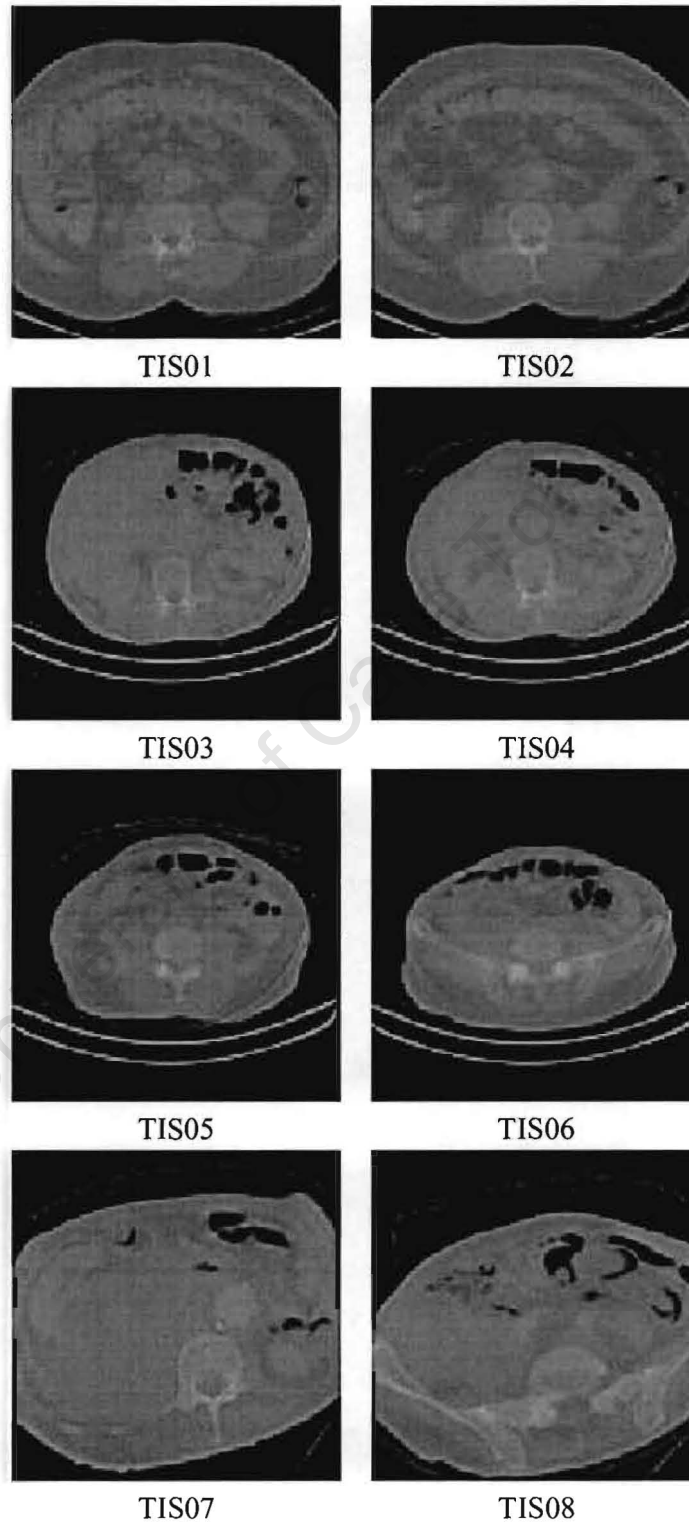
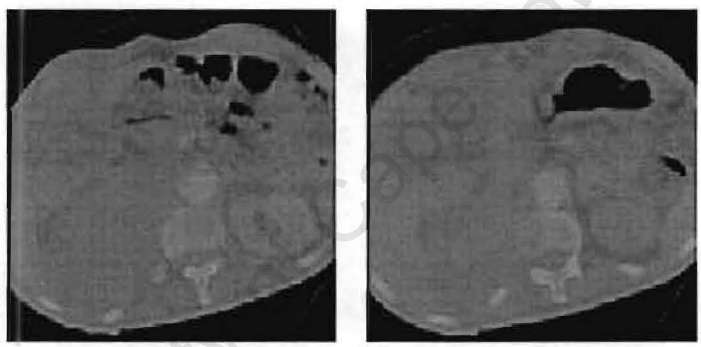


Figure A.3: Sample images in the TIS series



TIS09 TIS10

Figure A.4: Sample images in the TIS series

Appendix B

Experimental Results

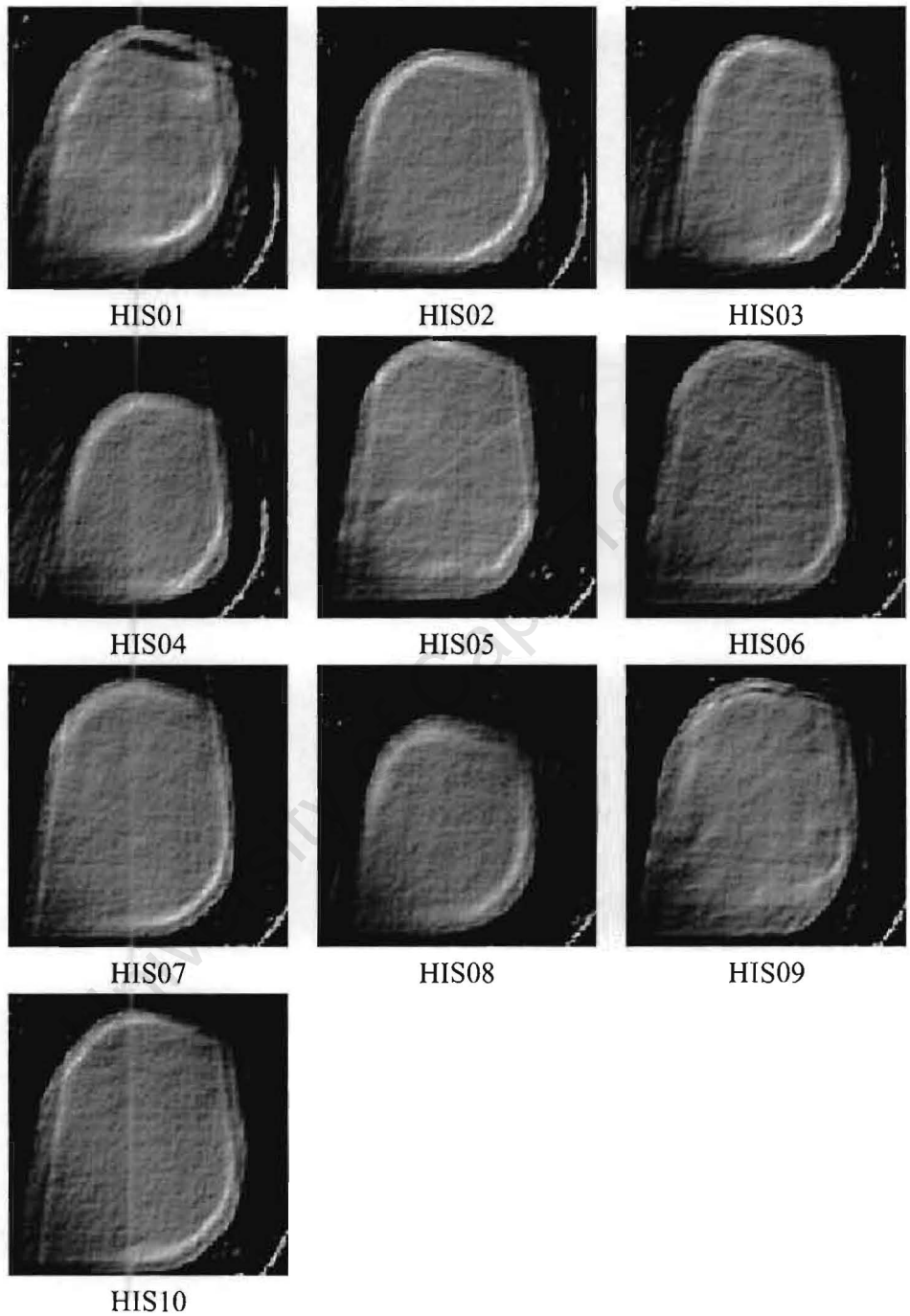


Figure B.1: Experiment 1. Maximum likelihood reconstruction using the Convex algorithm.

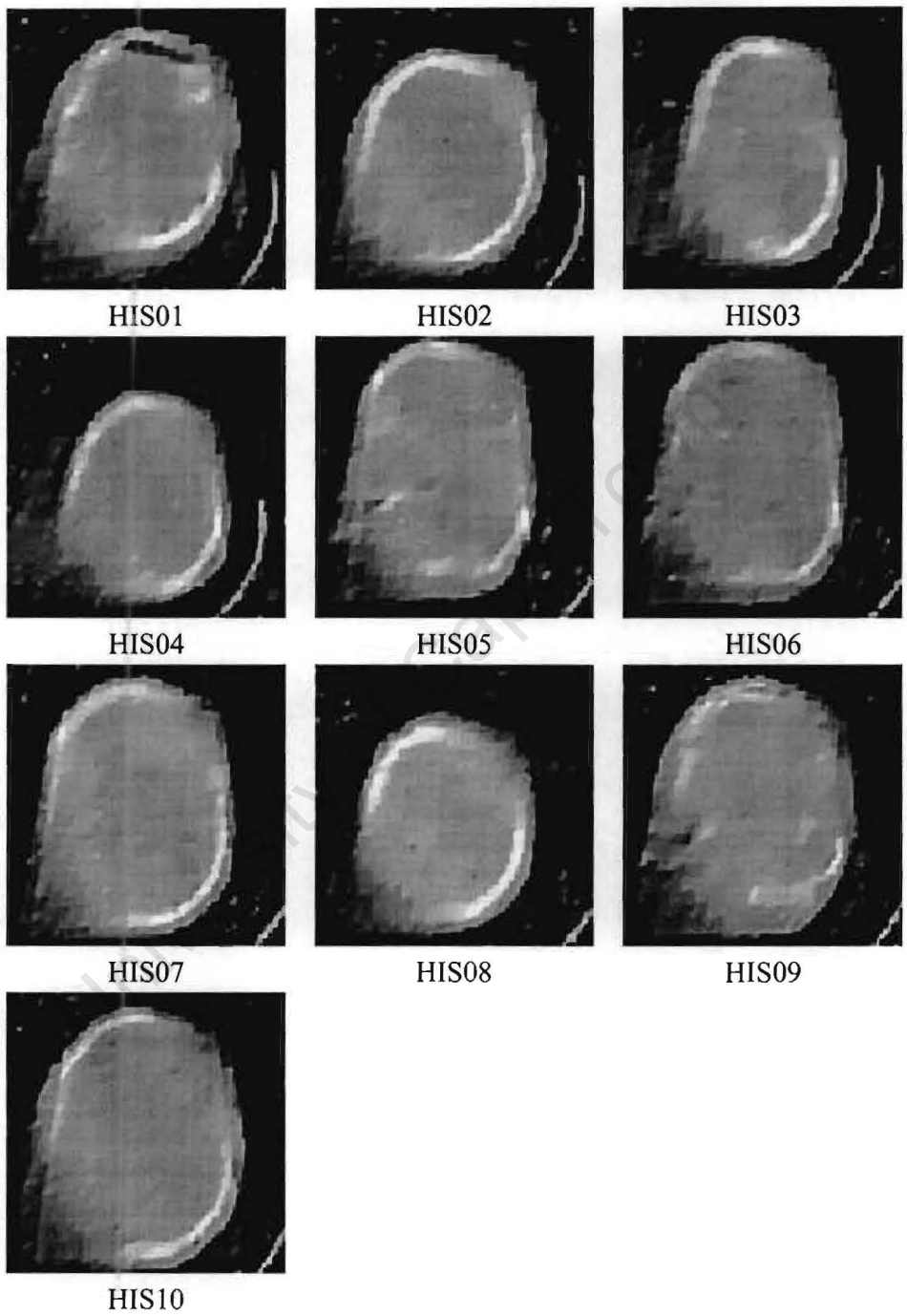


Figure B.2: Experiment 1. MAP reconstruction using the Convex algorithm.

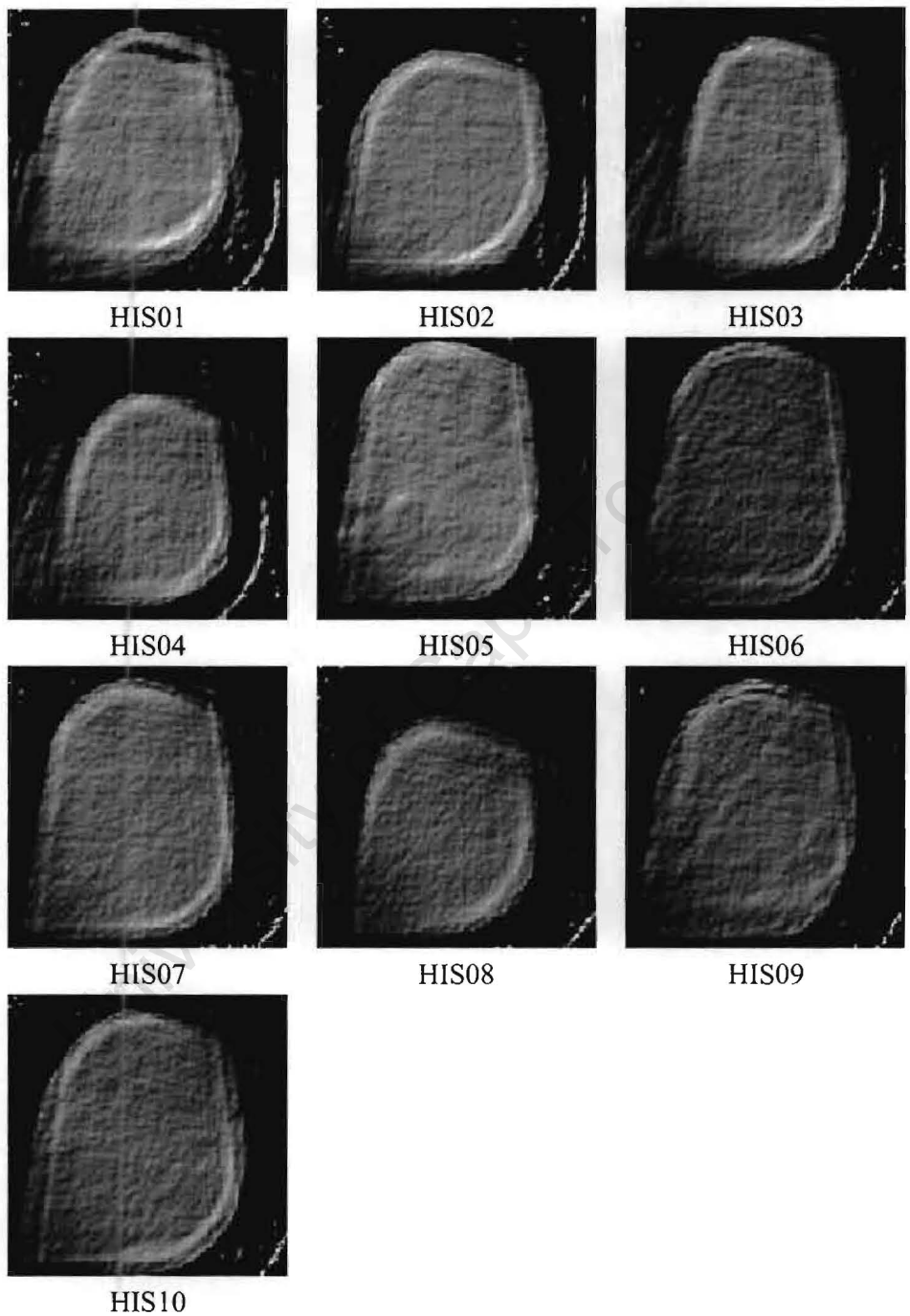


Figure B.3: Experiment 2. Maximum likelihood reconstruction using the Convex algorithm.

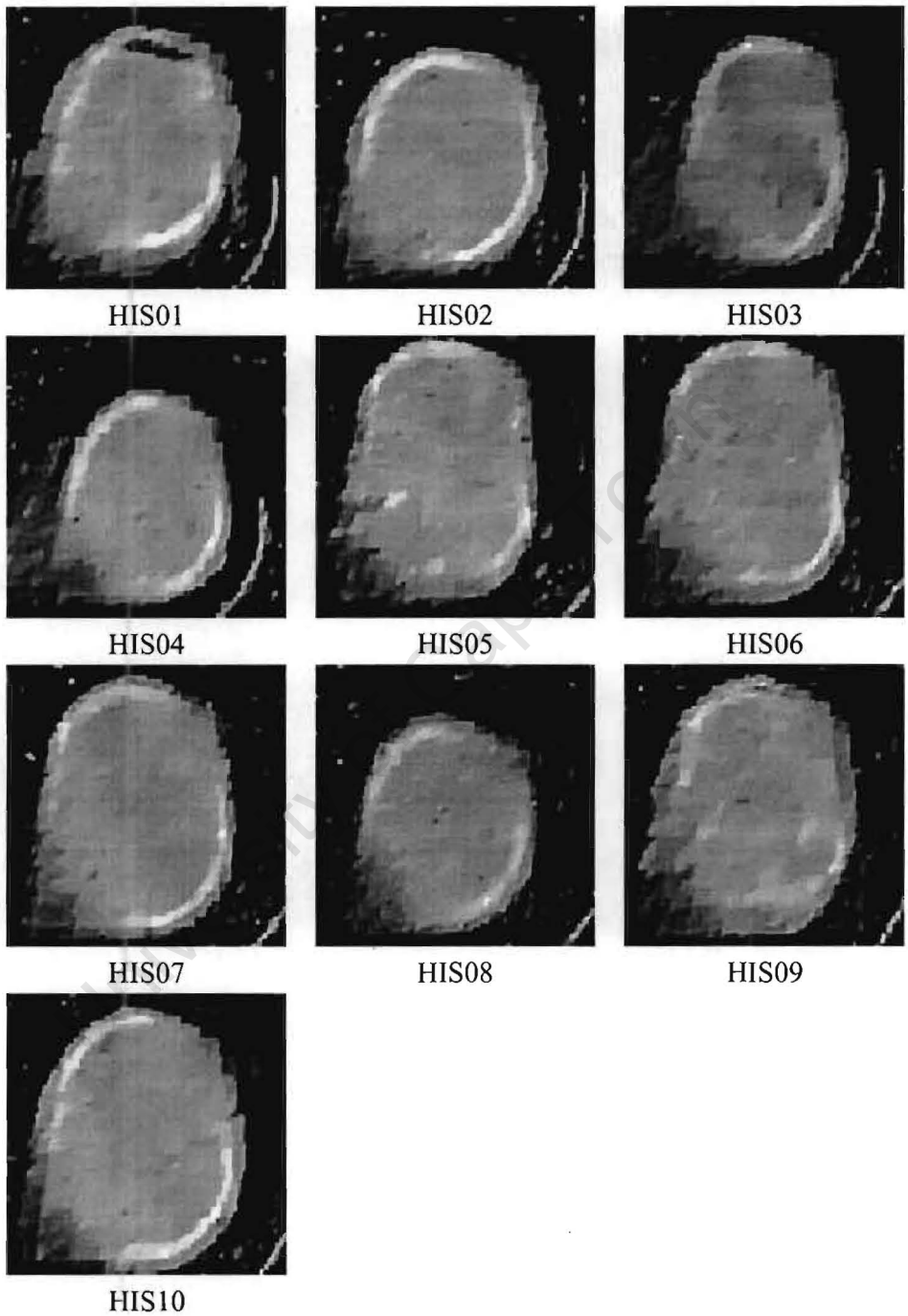


Figure B.4: Experiment 2. MAP reconstruction using the Convex algorithm.

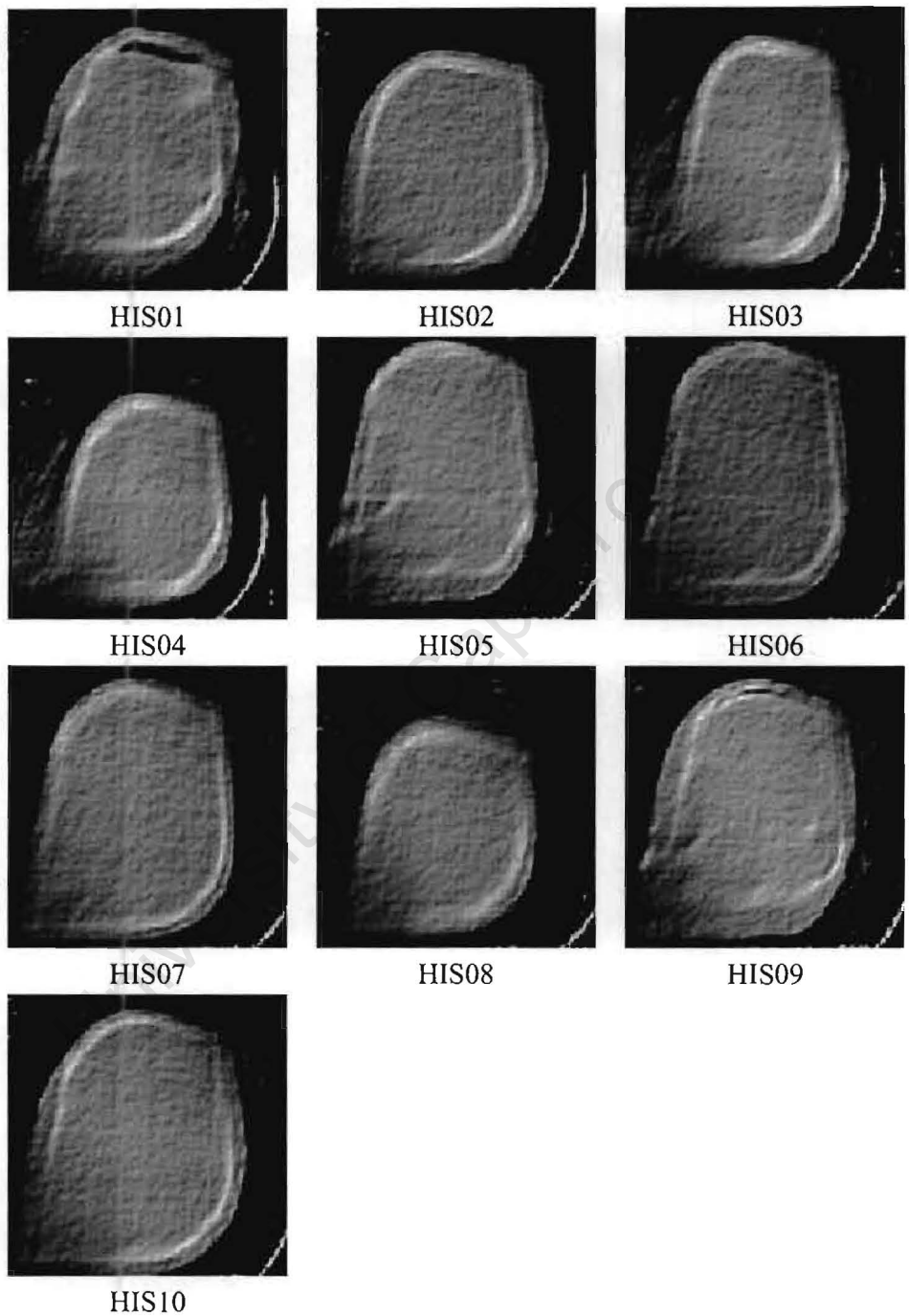


Figure B.5: Experiment 3. Maximum likelihood reconstruction using the Convex algorithm.

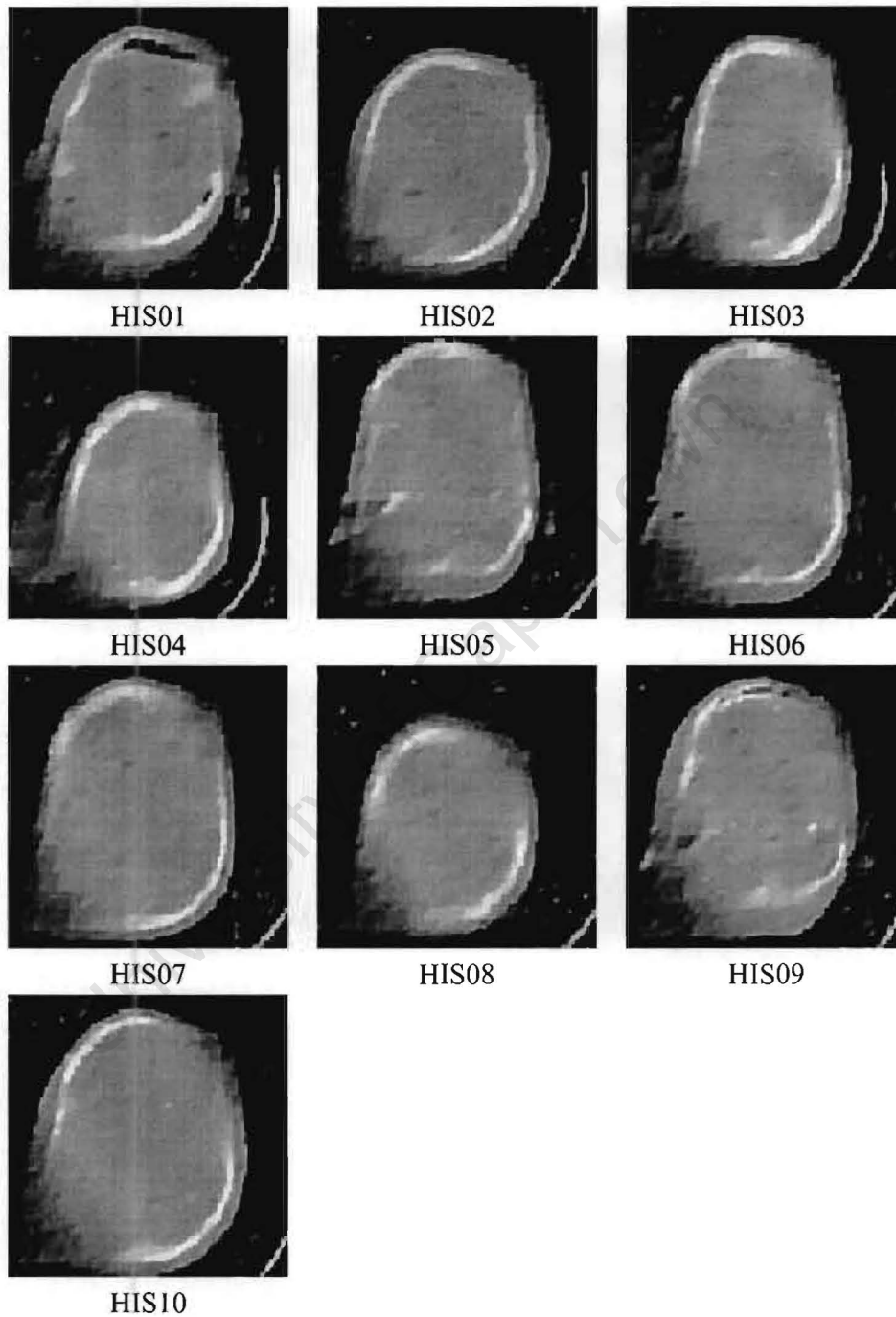


Figure B.6: Experiment 3. MAP reconstruction using the Convex algorithm.

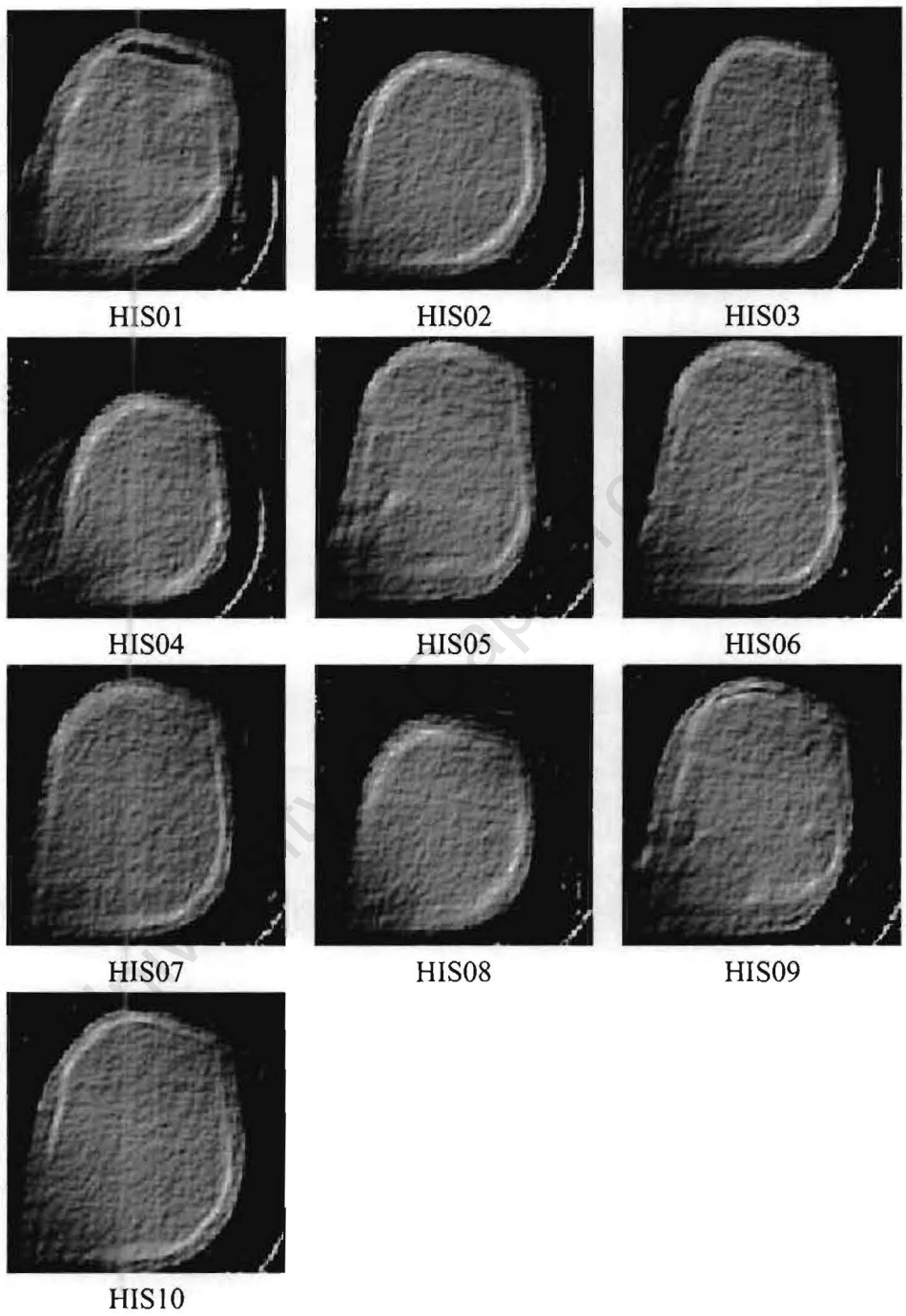


Figure B.7: Experiment 4. Maximum likelihood reconstruction using the Convex algorithm.

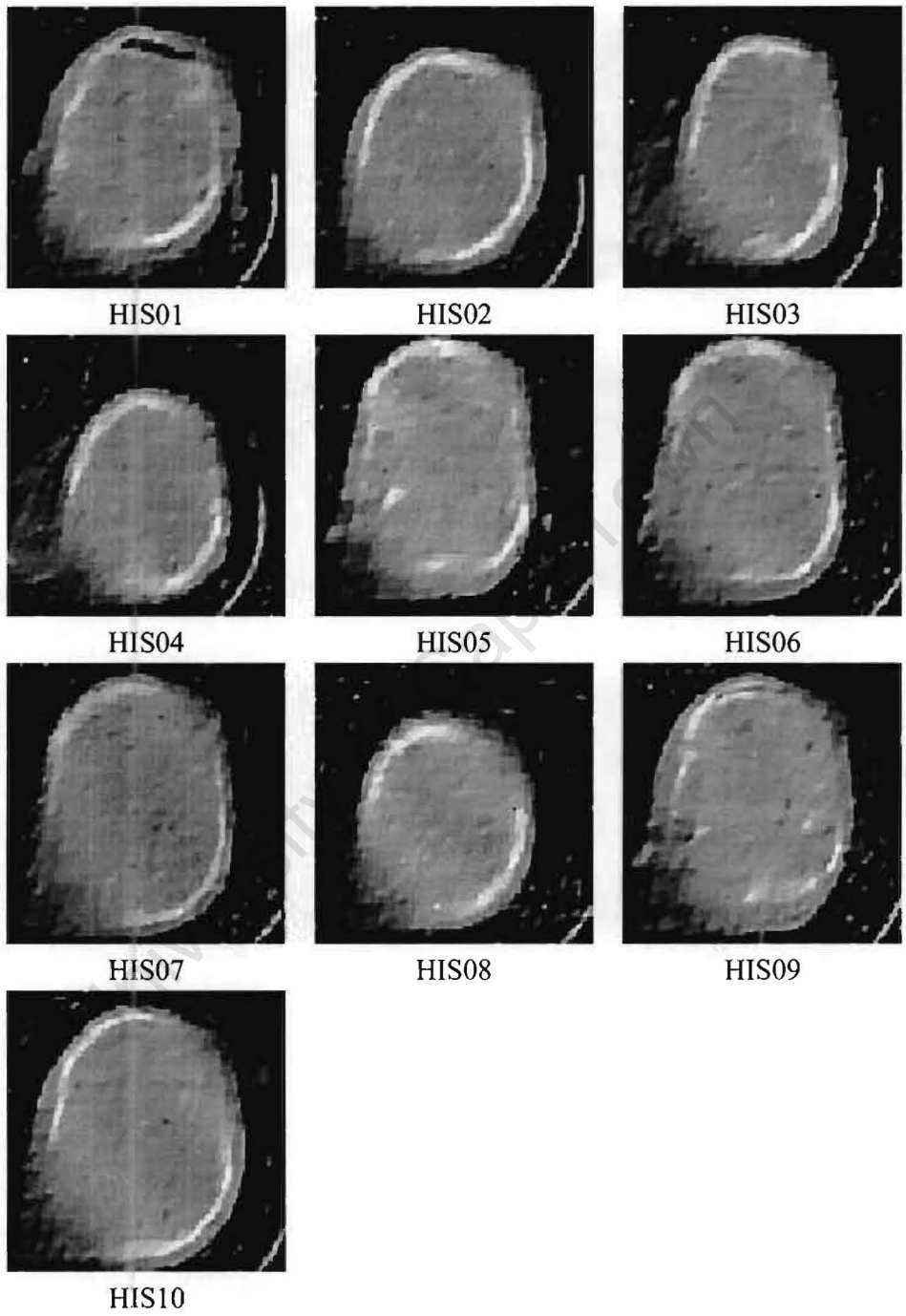


Figure B.8: Experiment 4. MAP reconstruction using the Convex algorithm.

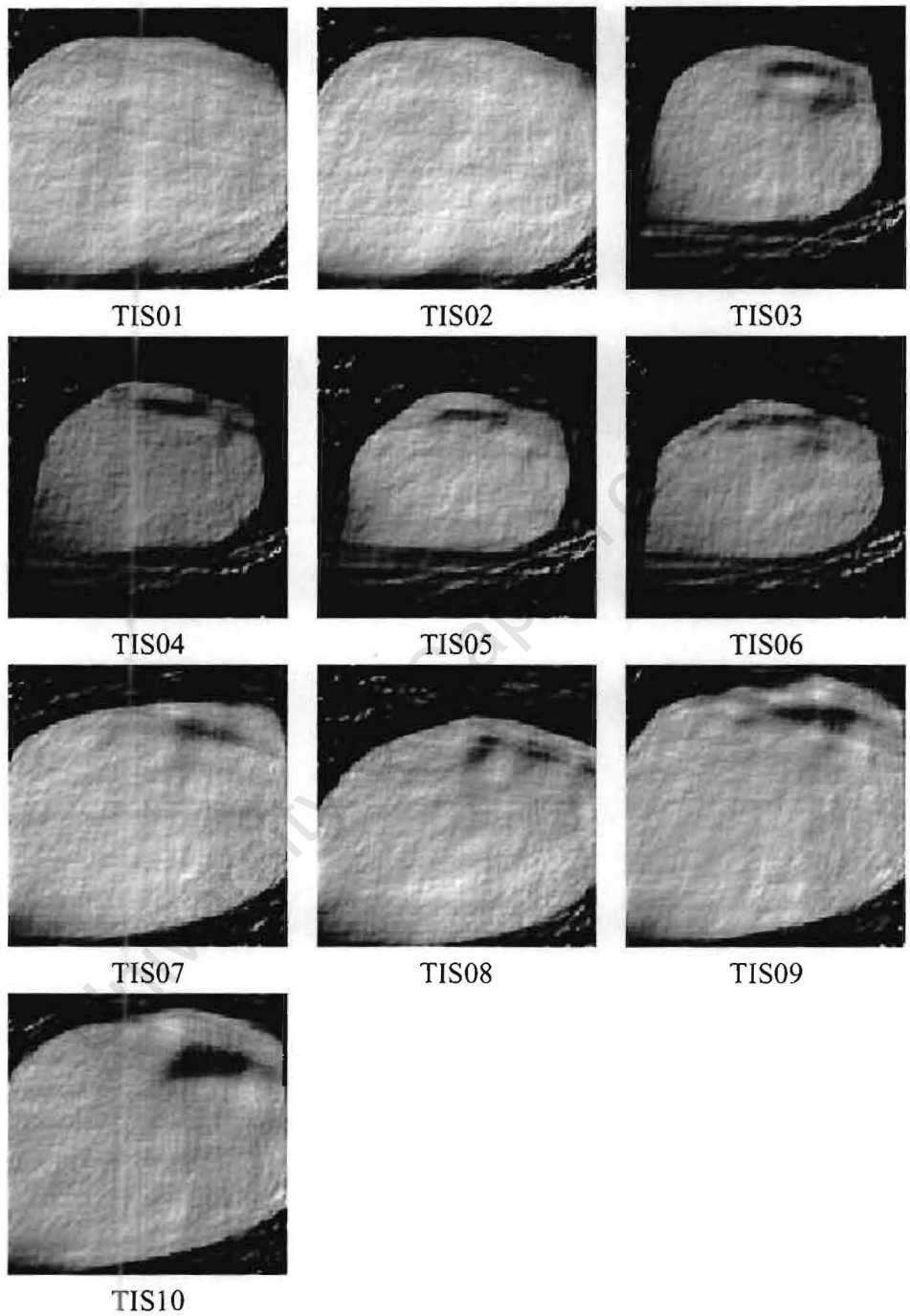


Figure B.9: Experiment 5. Maximum likelihood reconstruction using the Convex algorithm.

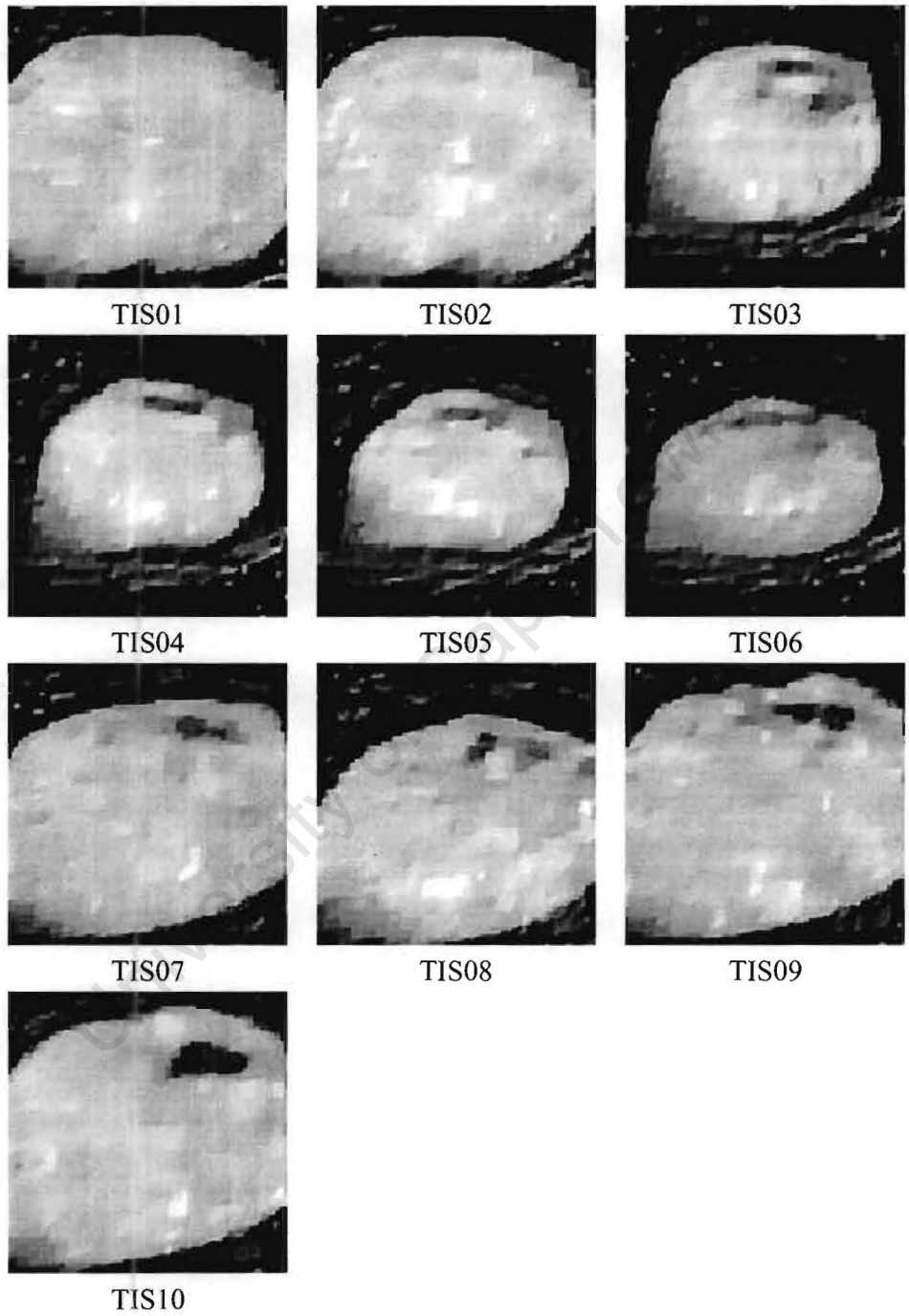


Figure B.10: Experiment 5. MAP reconstruction using the Convex algorithm.

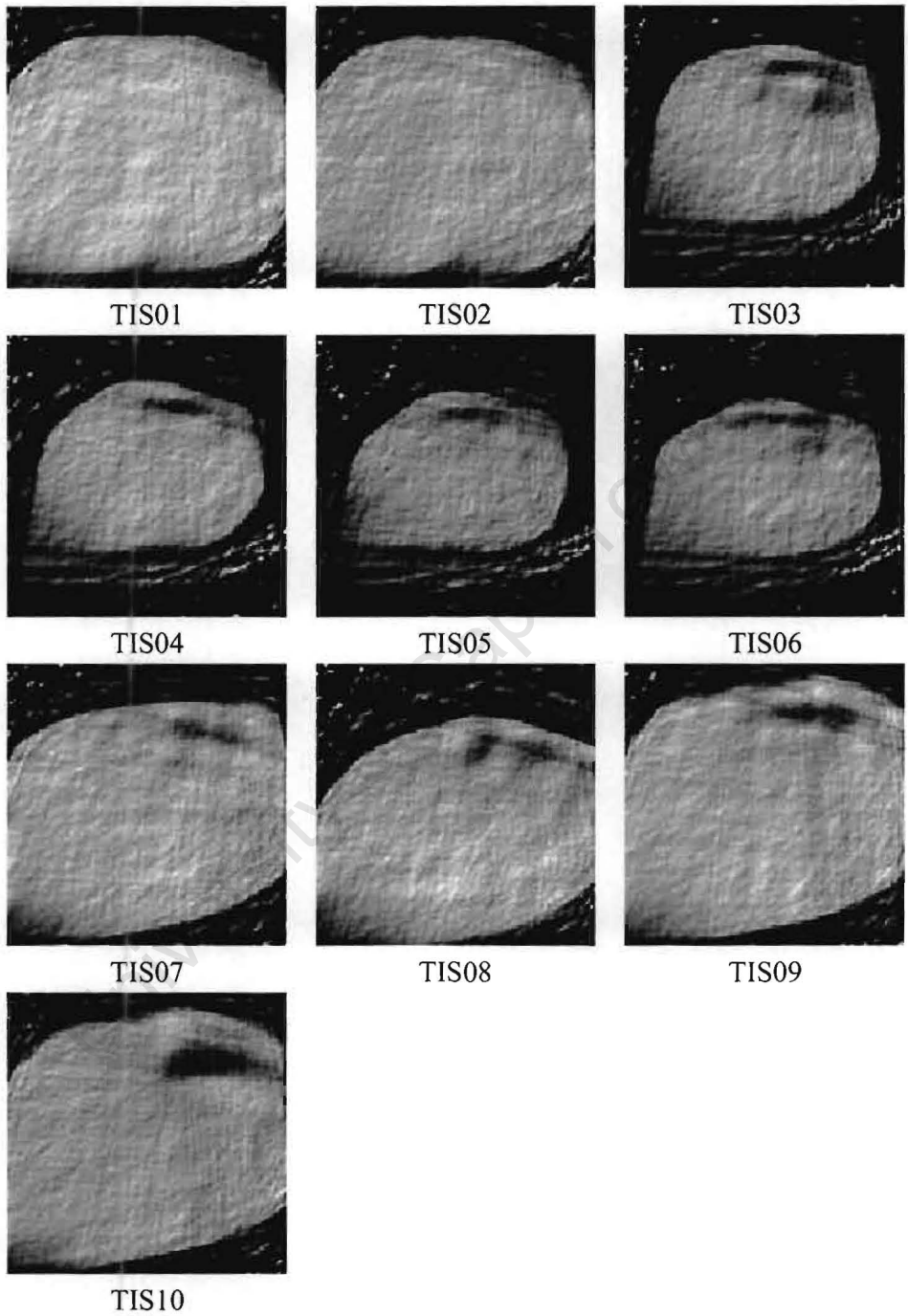


Figure B.11: Experiment 6. Maximum likelihood reconstruction using the Convex algorithm.

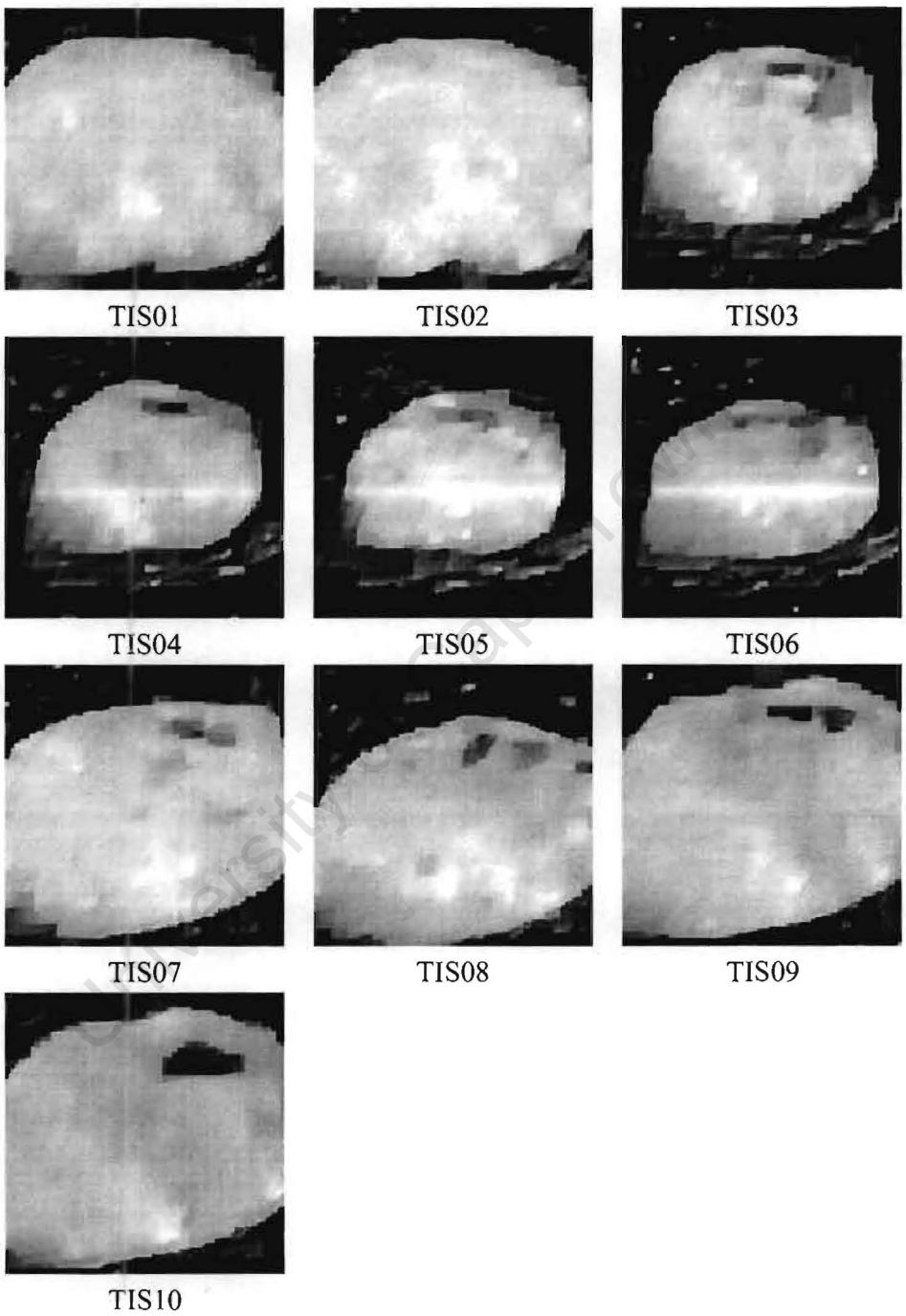


Figure B.12: Experiment 6. MAP reconstruction using the Convex algorithm.

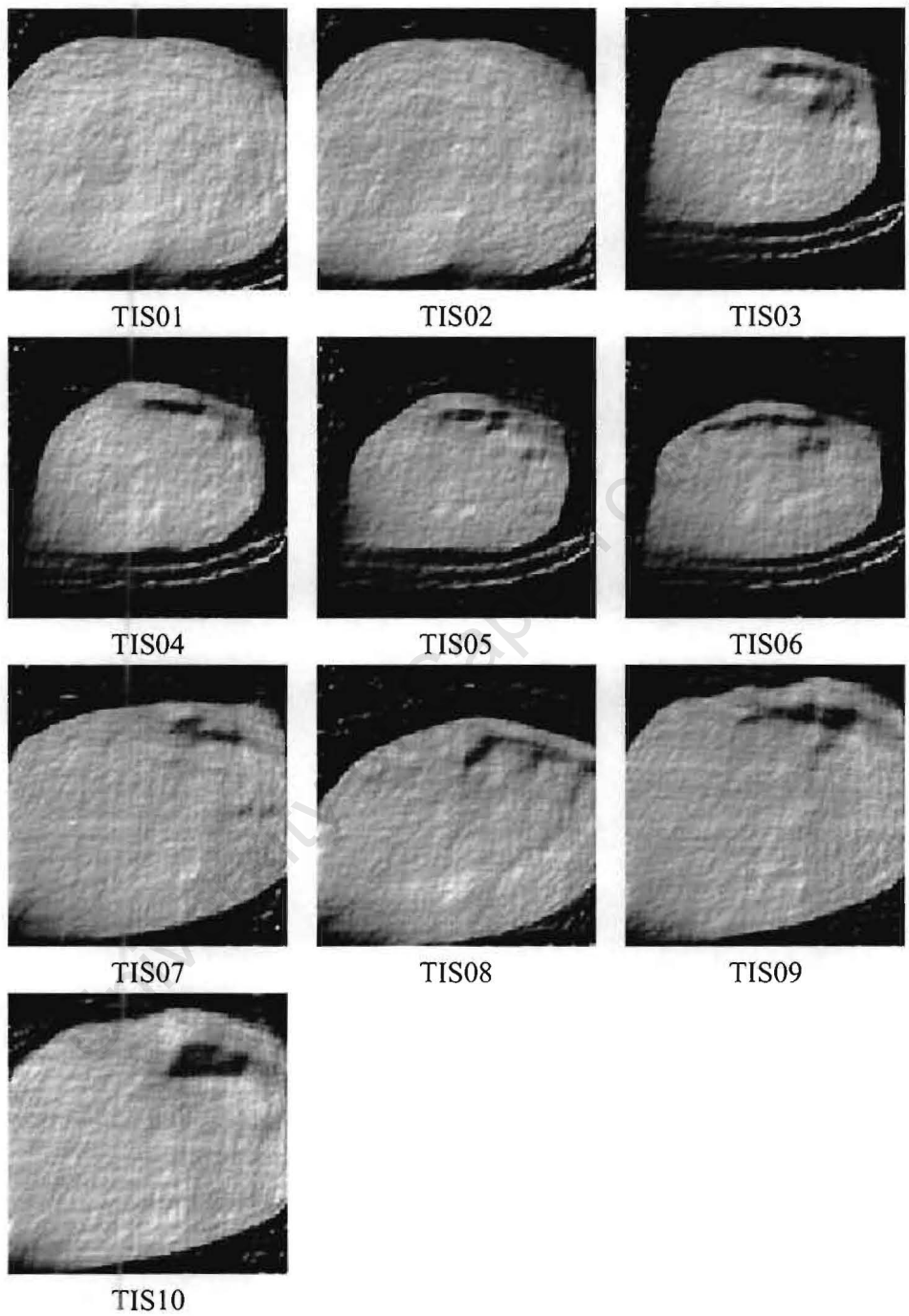


Figure B.13: Experiment 7. Maximum likelihood reconstruction using the Convex algorithm.

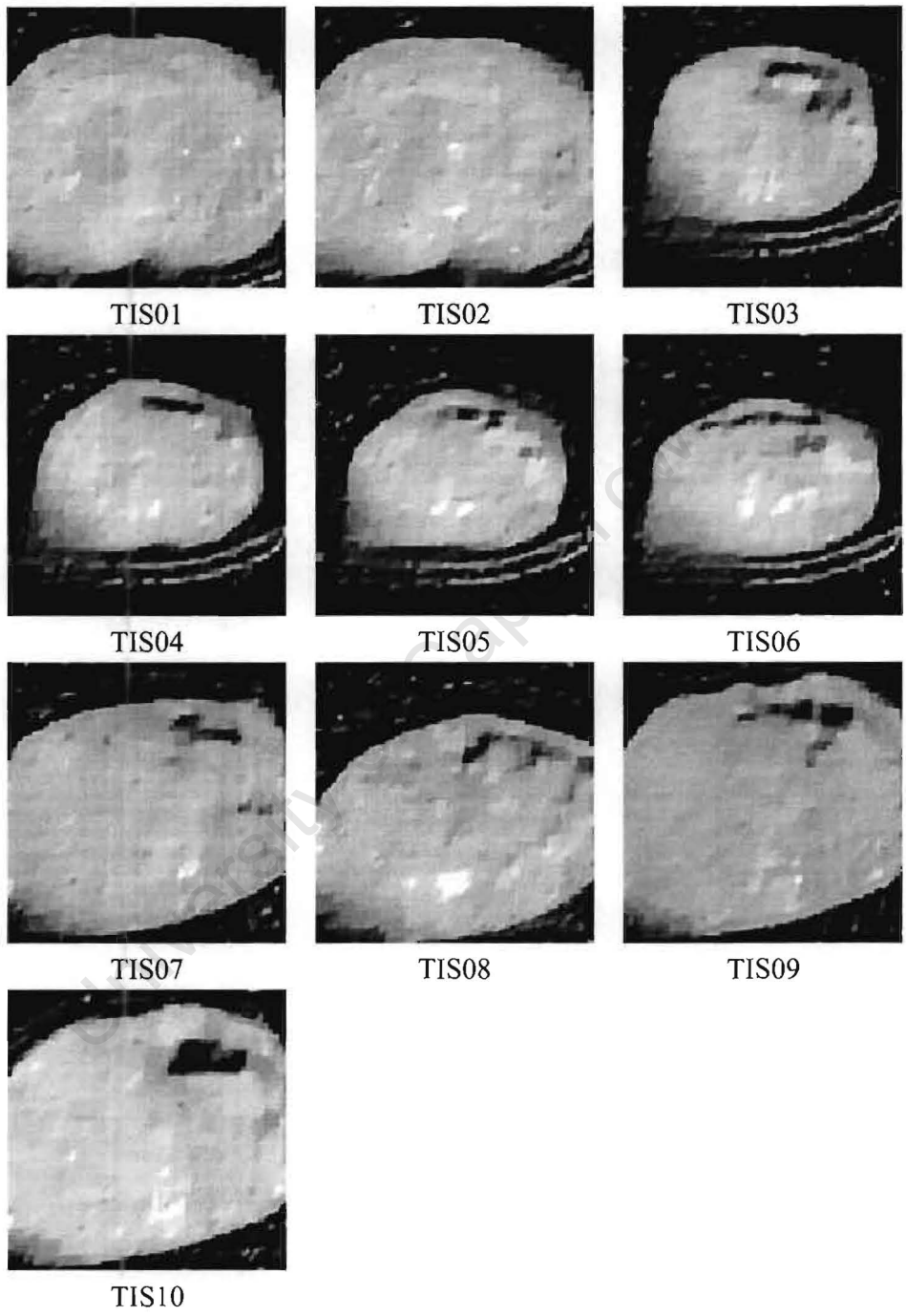


Figure B.14: Experiment 7. MAP reconstruction using the Convex algorithm.

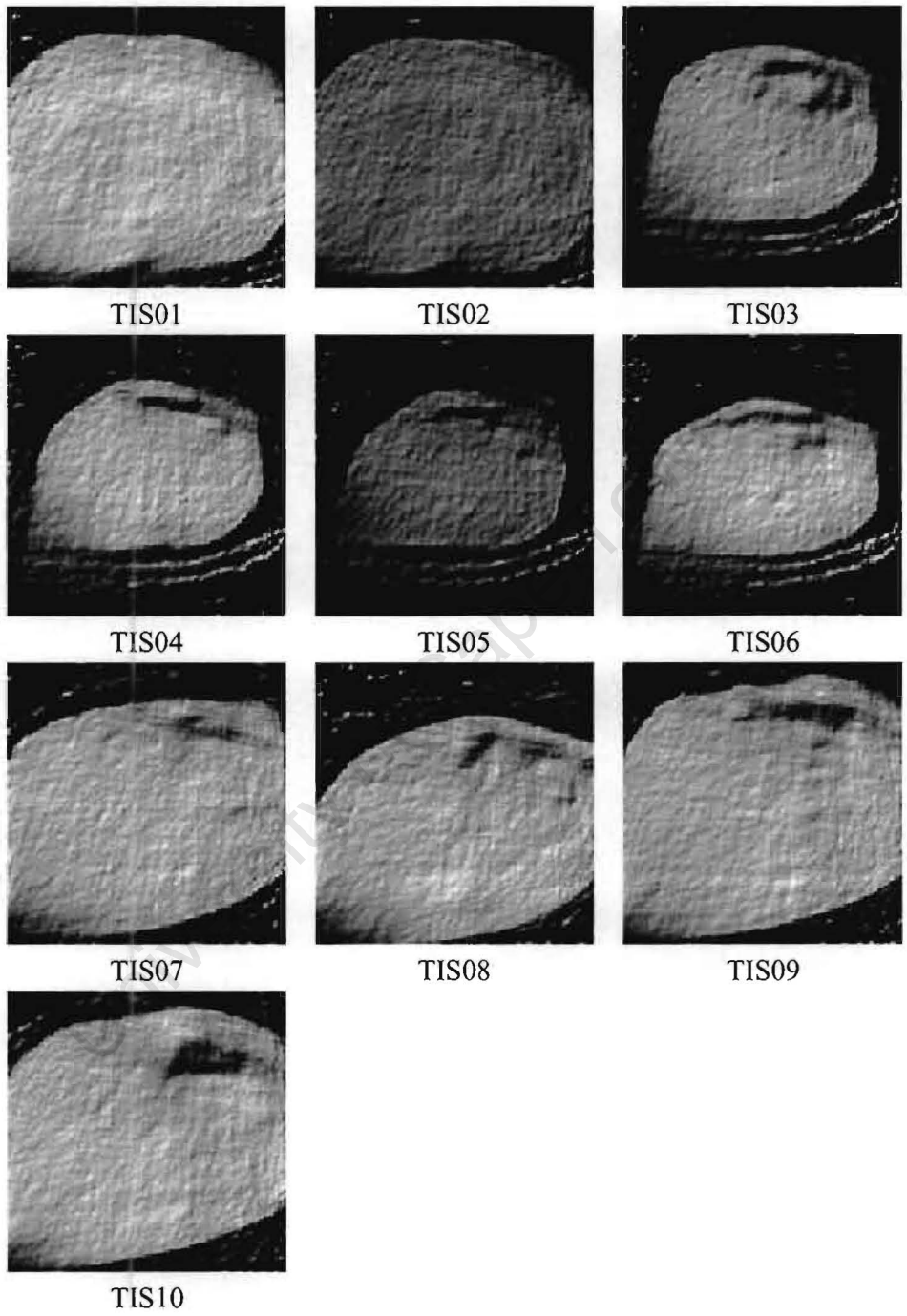


Figure B.15: Experiment 8. Maximum likelihood reconstruction using the Convex algorithm.

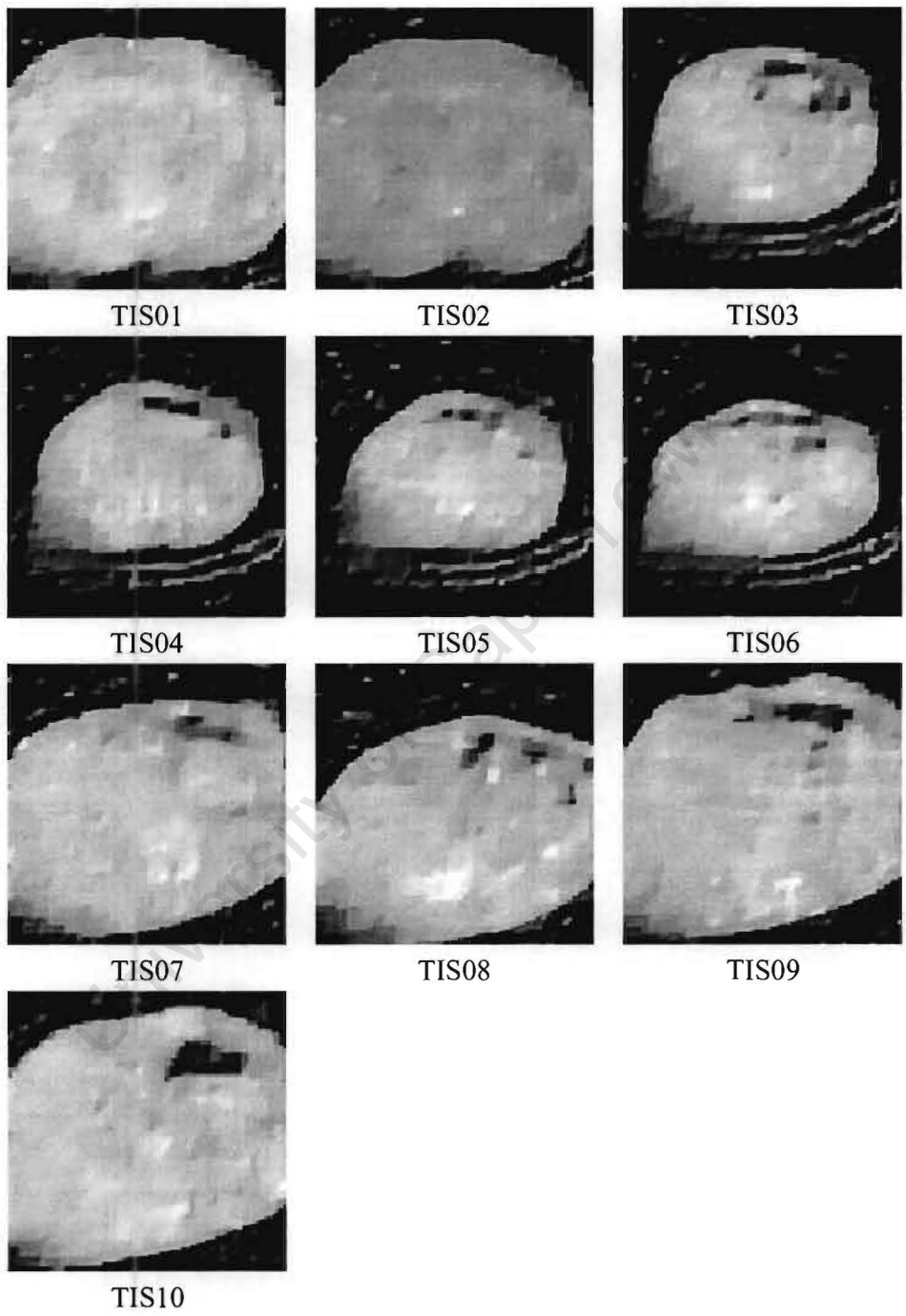


Figure B.16: Experiment 8. MAP reconstruction using the Convex algorithm.

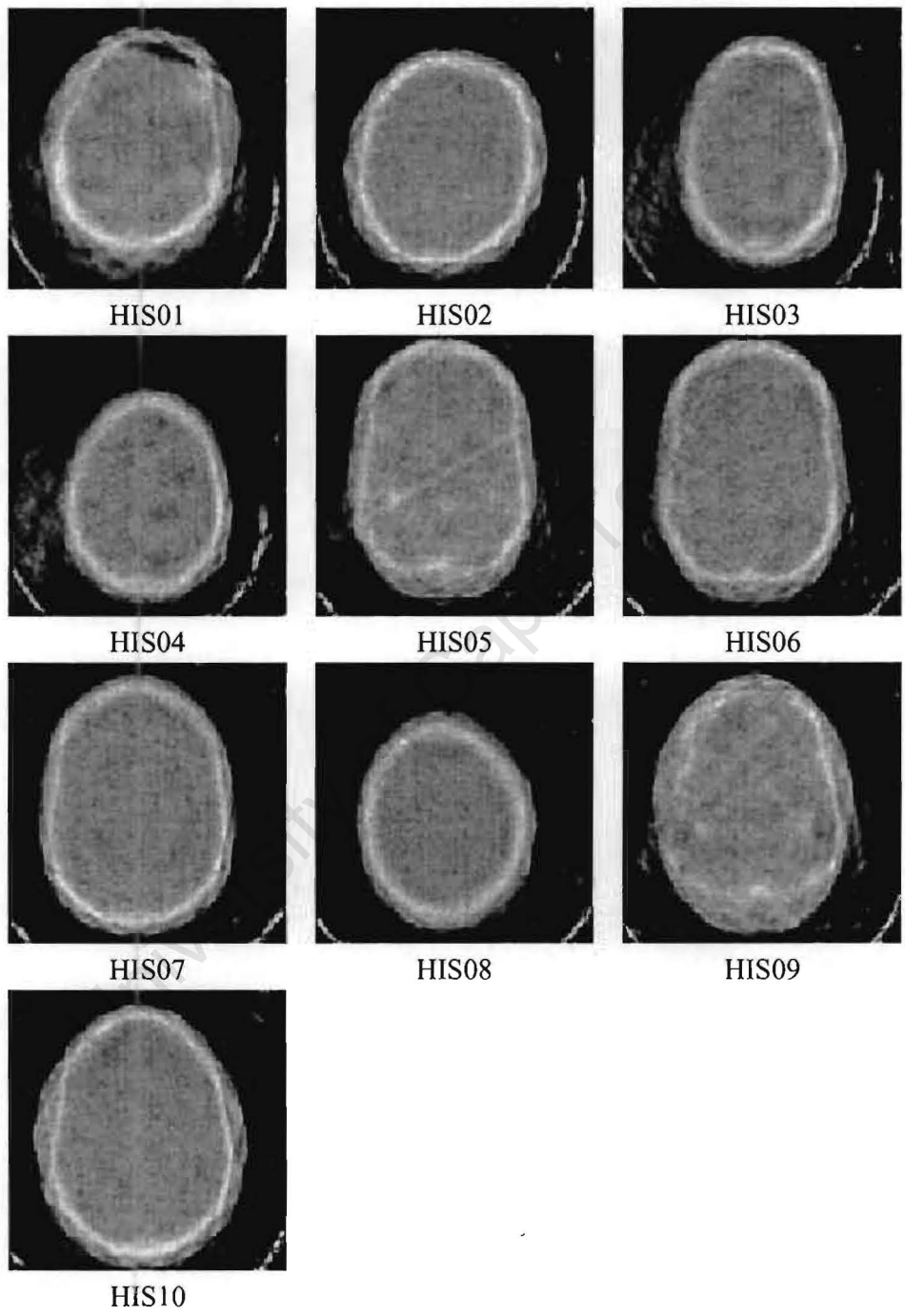


Figure B.17: Experiment 9. Maximum likelihood reconstruction using the Convex algorithm.

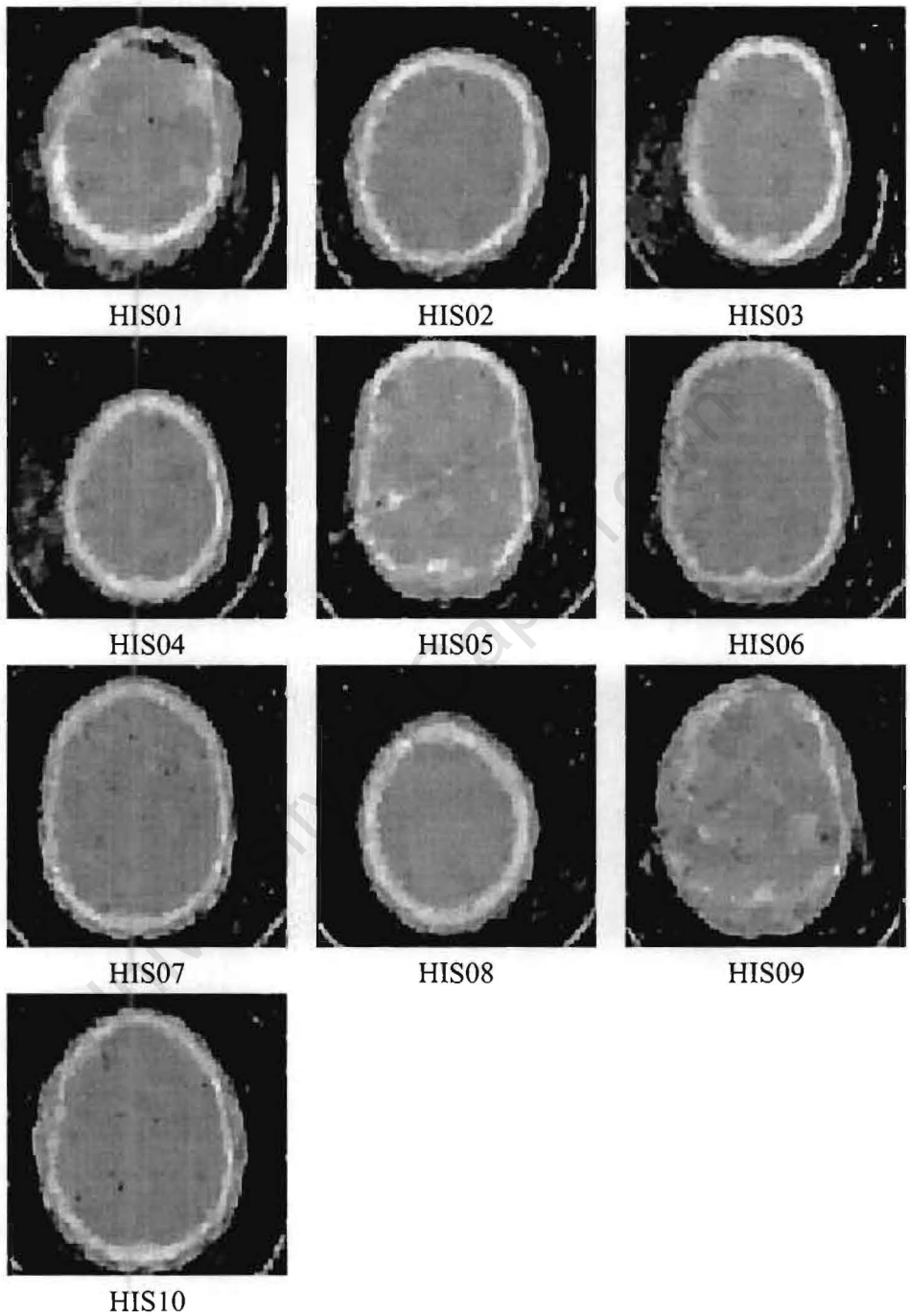


Figure B.18: Experiment 9. MAP reconstruction using the Convex algorithm.

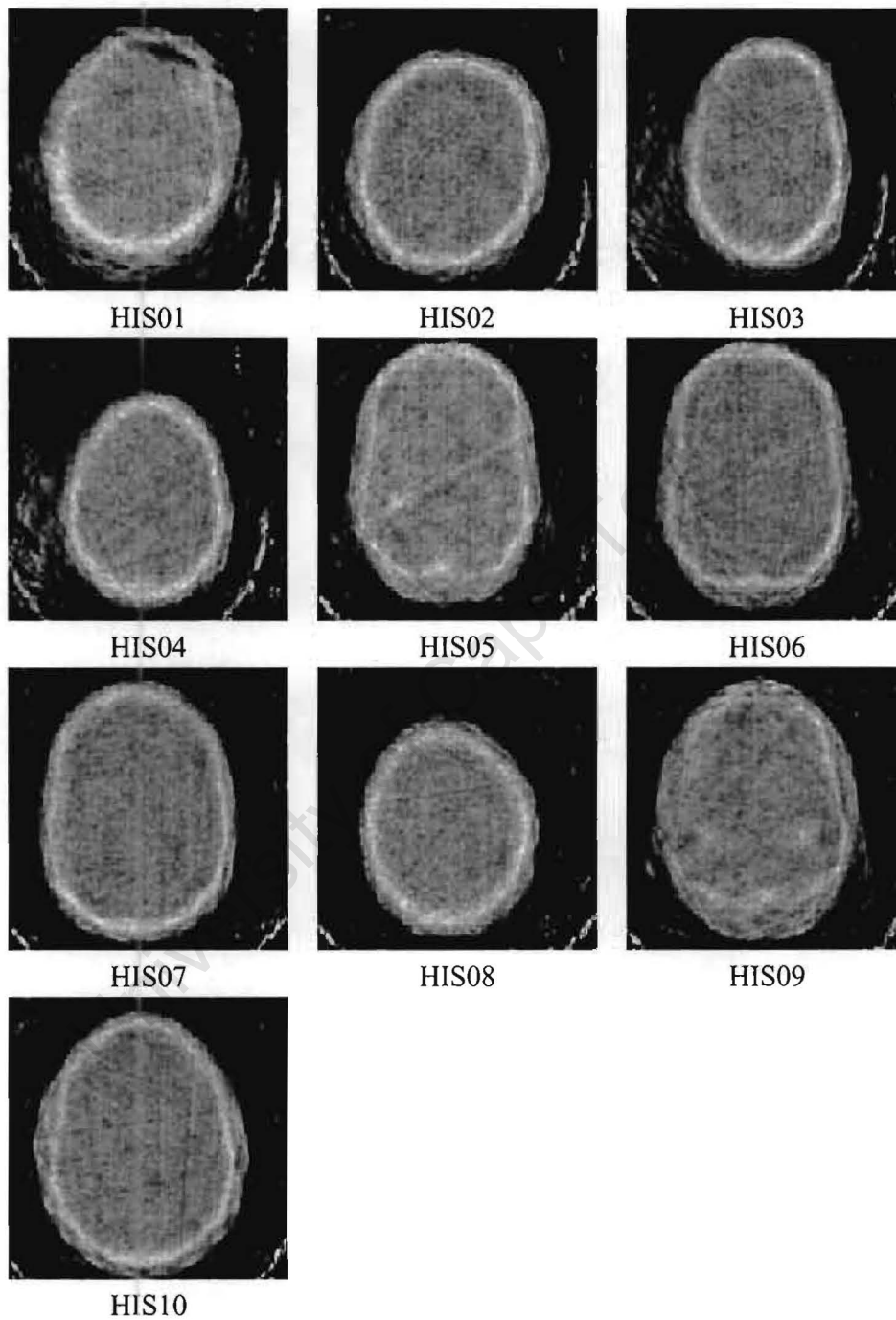


Figure B.19: Experiment 10. Maximum likelihood reconstruction using the Convex algorithm.

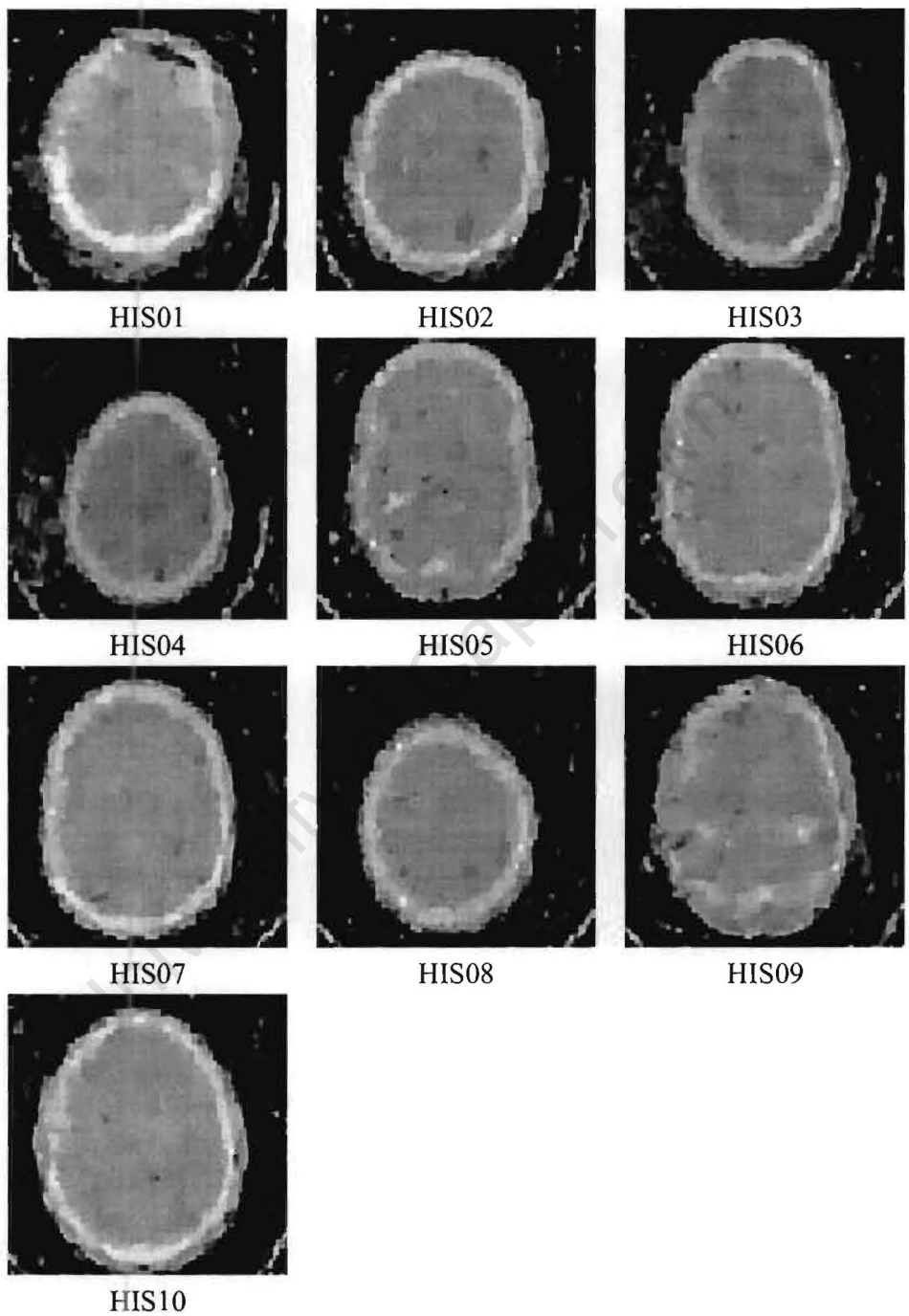


Figure B.20: Experiment 10. MAP reconstruction using the Convex algorithm.

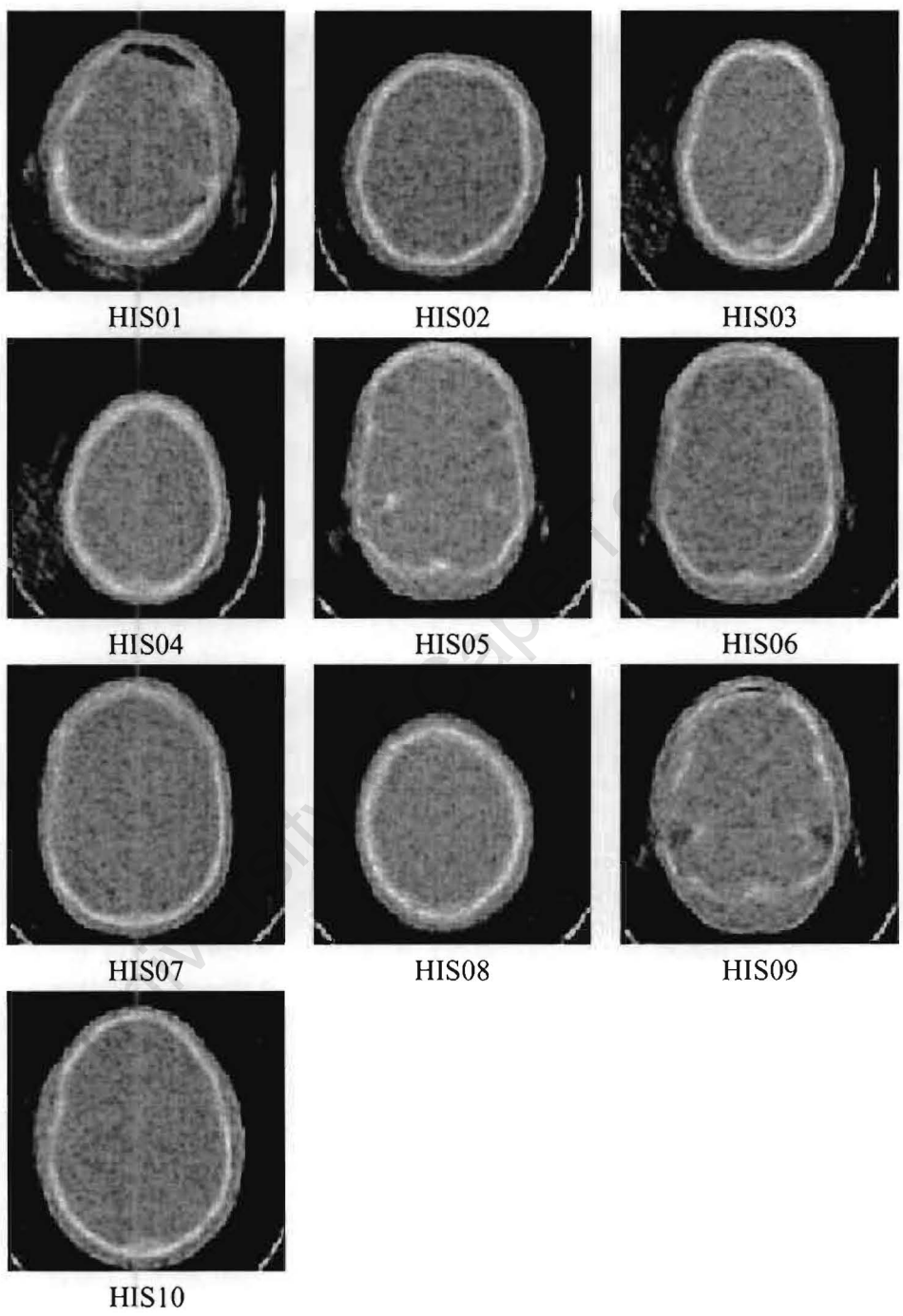


Figure B.21: Experiment 11. Maximum likelihood reconstruction using the Convex algorithm.

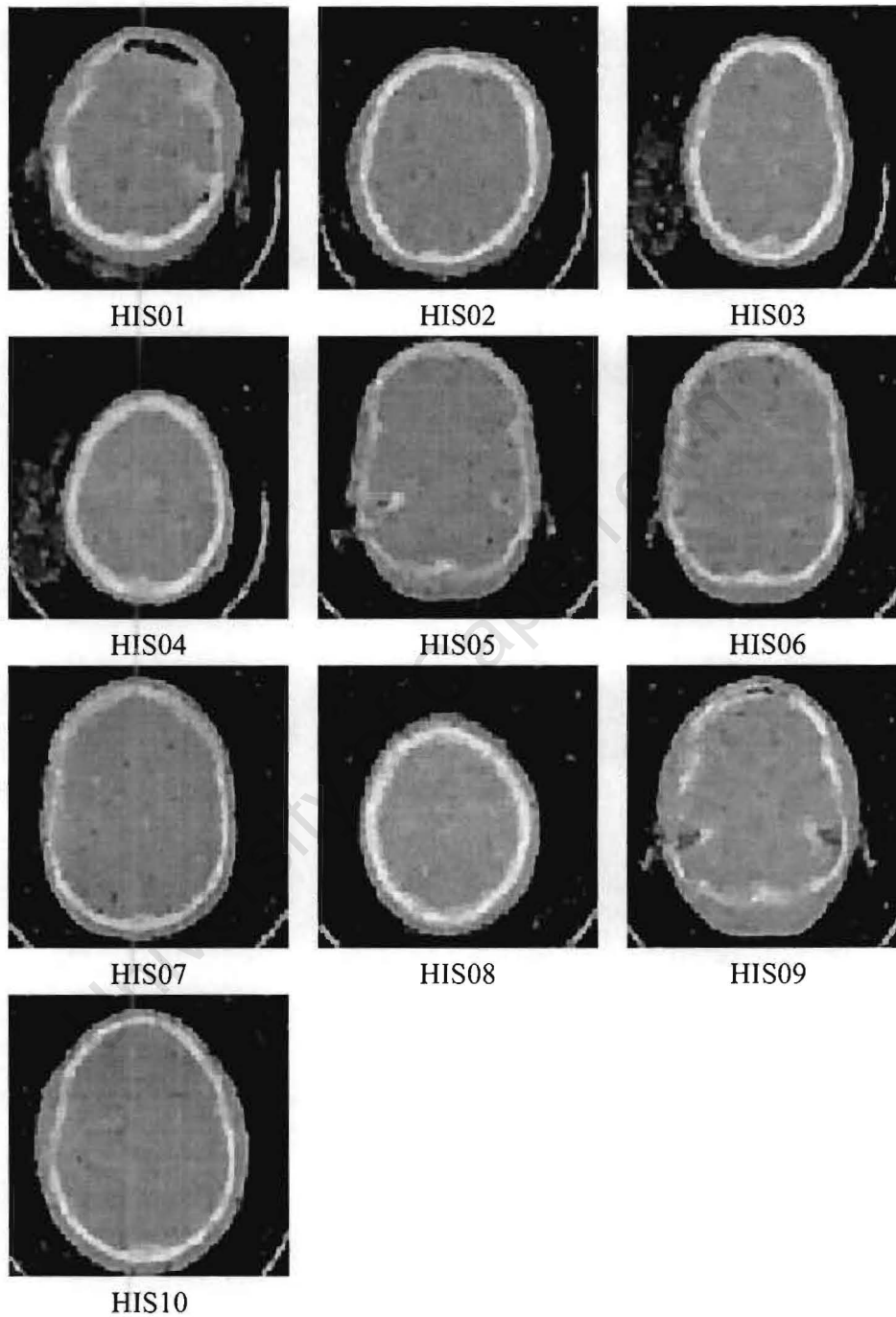


Figure B.22: Experiment 11. MAP reconstruction using the Convex algorithm.

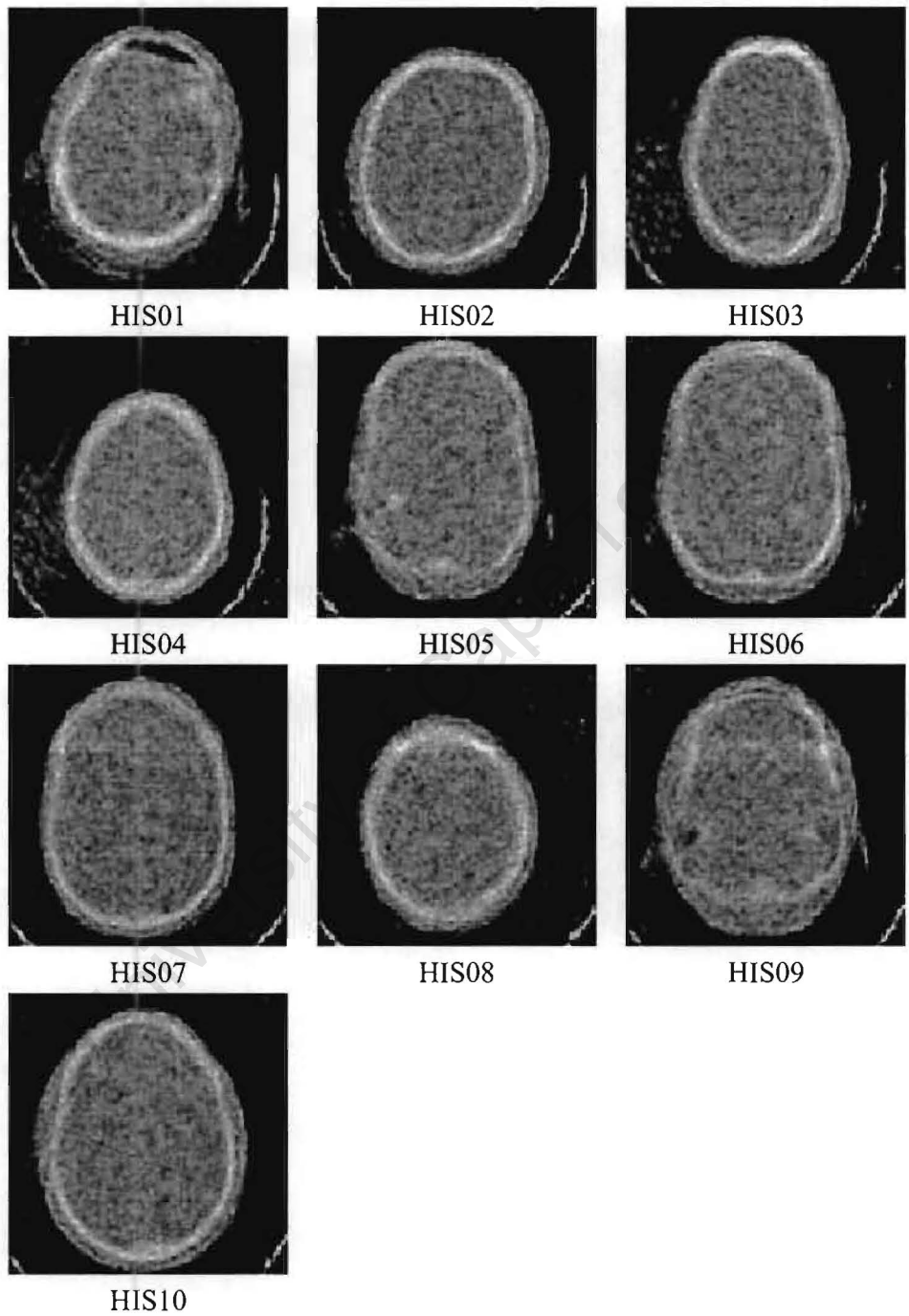


Figure B.23: Experiment 12. Maximum likelihood reconstruction using the Convex algorithm.

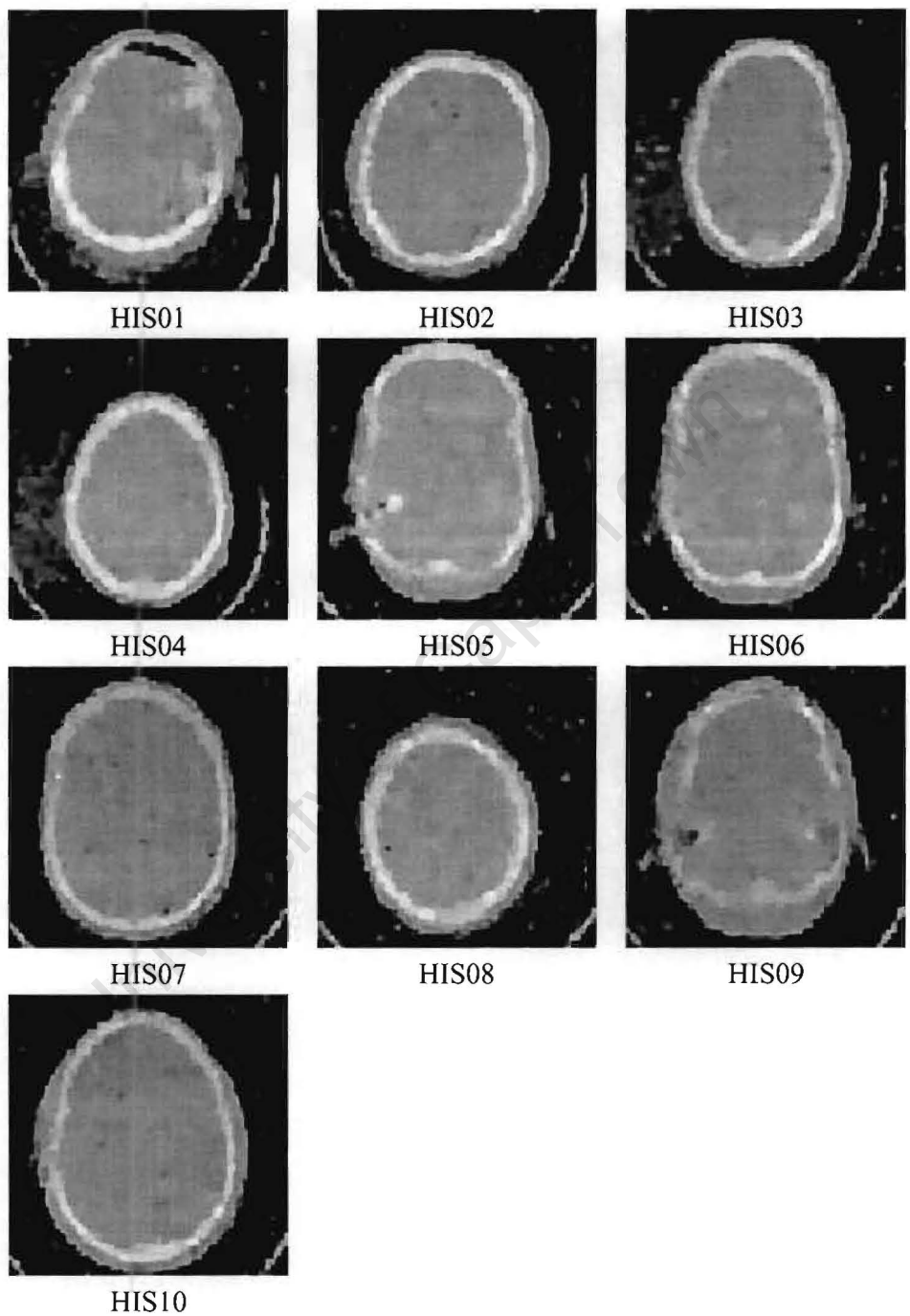


Figure B.24: Experiment 12. MAP reconstruction using the Convex algorithm.

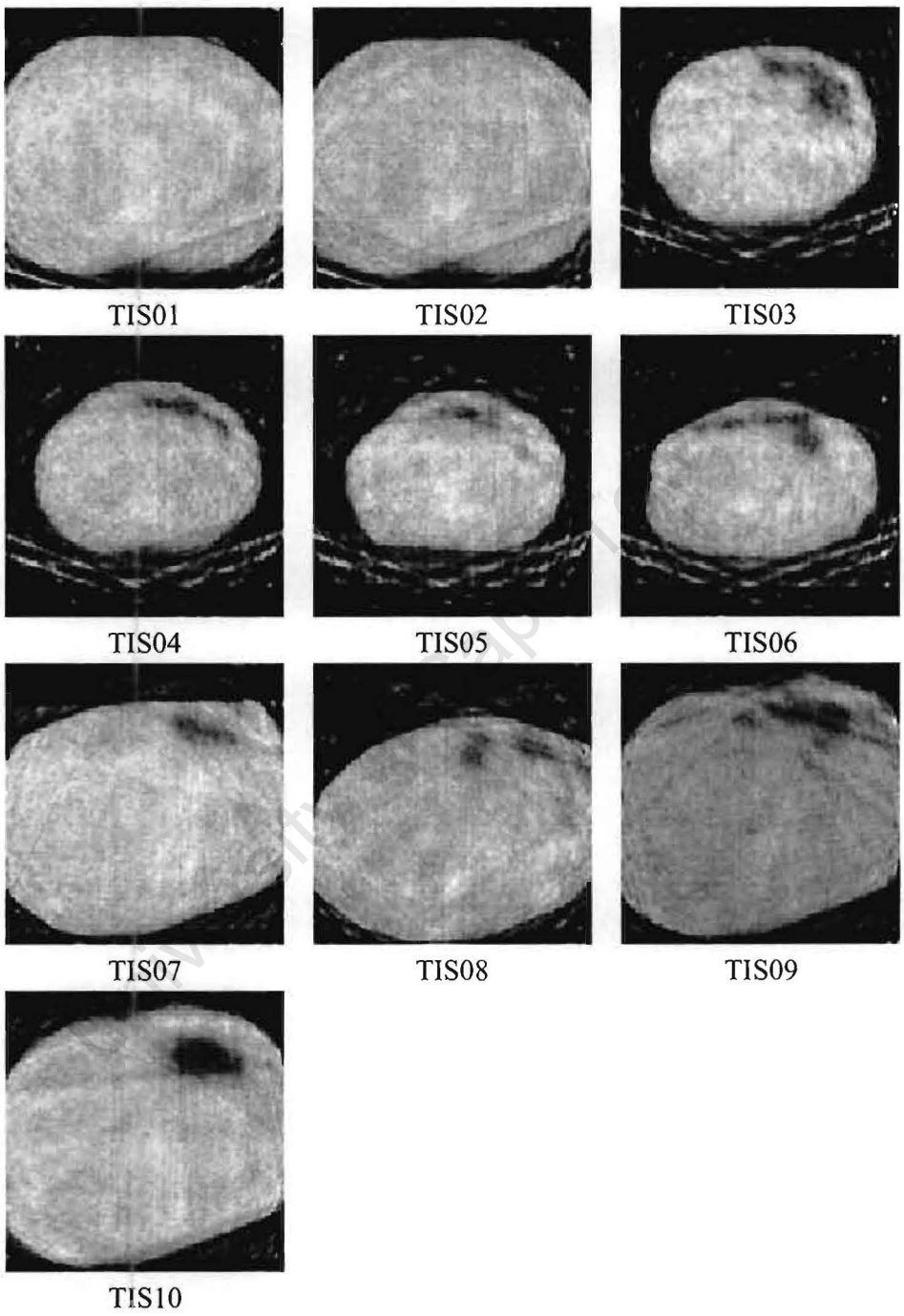


Figure B.25: Experiment 13. Maximum likelihood reconstruction using the Convex algorithm.

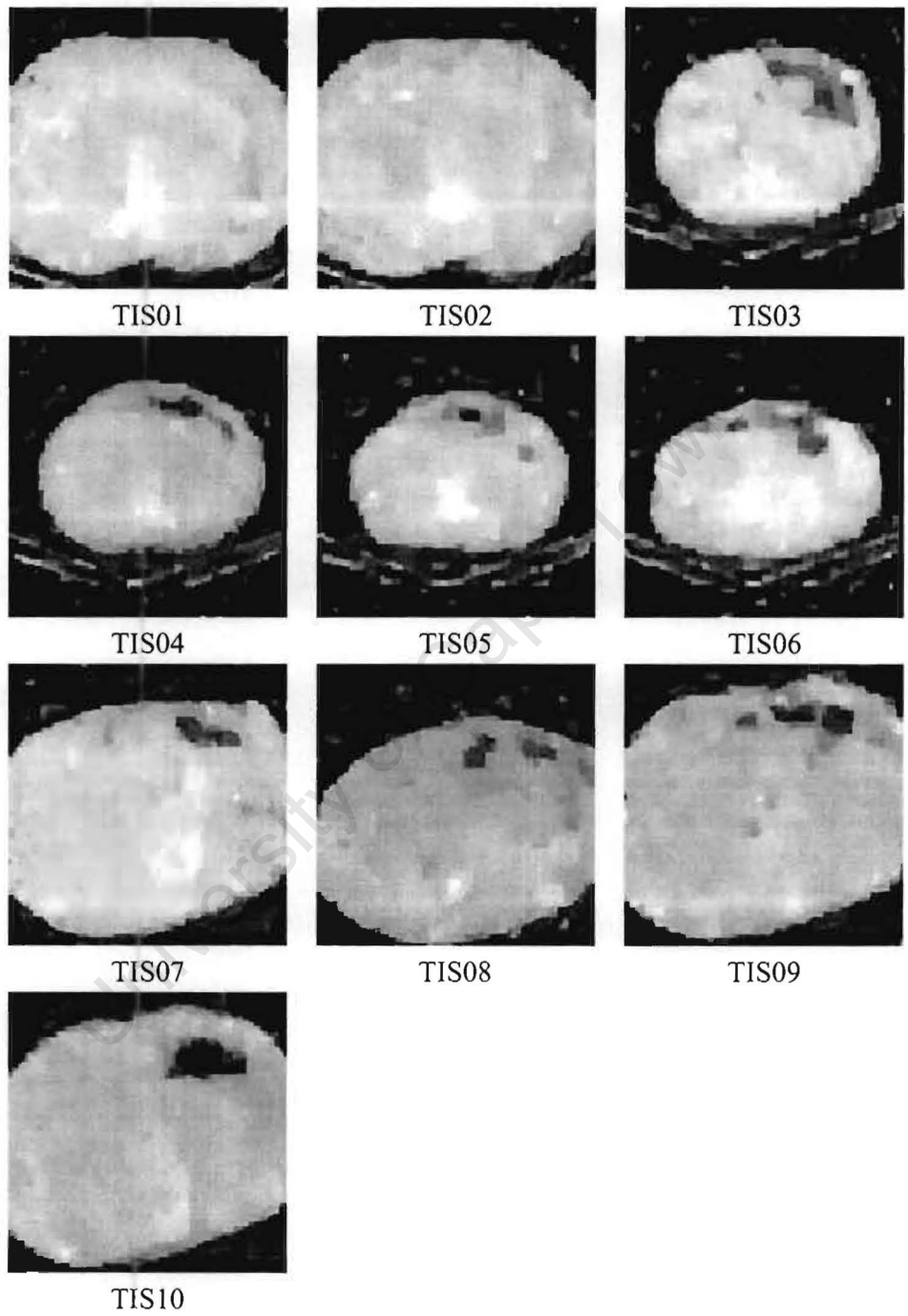


Figure B.26: Experiment 13. MAP reconstruction using the Convex algorithm.

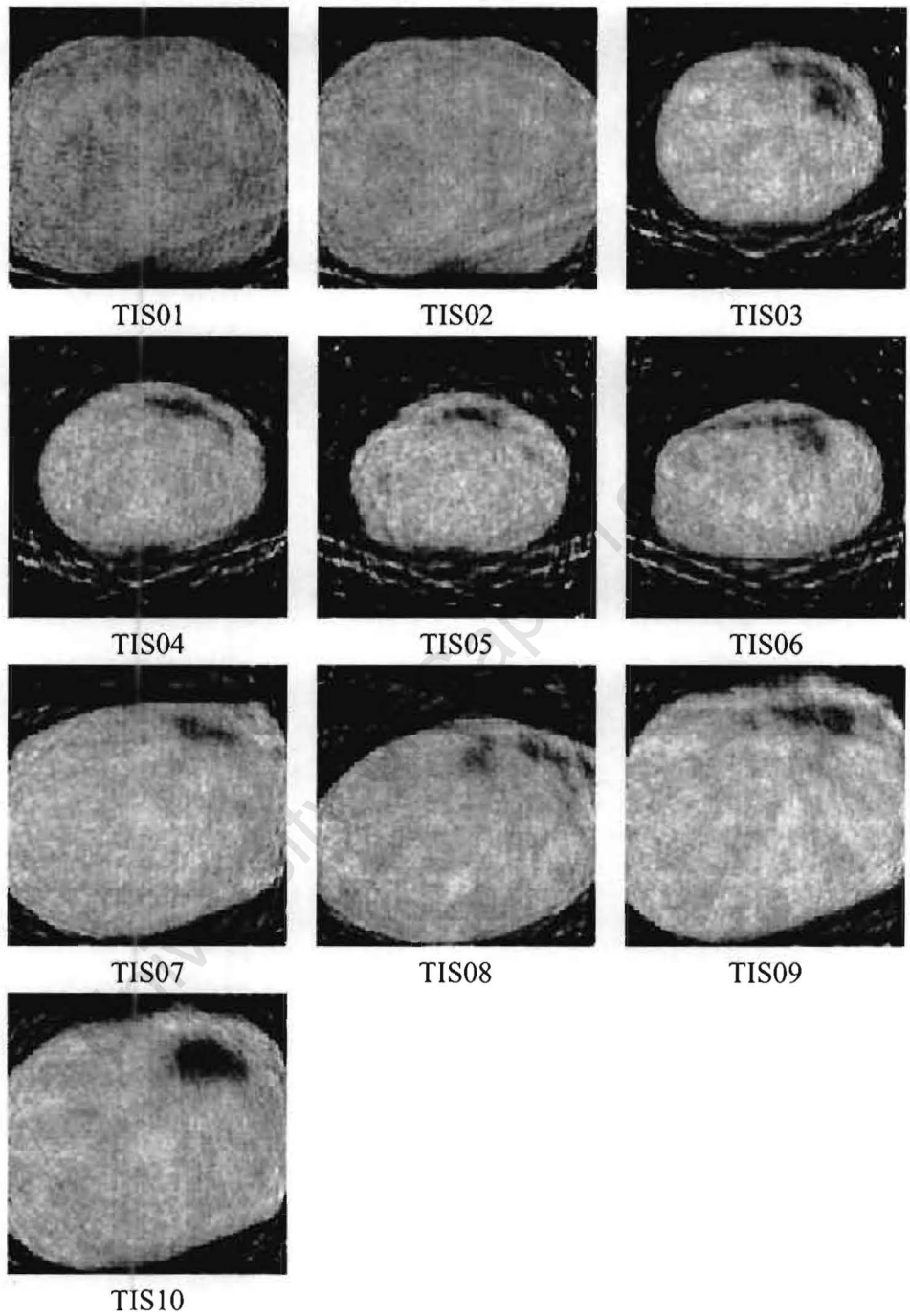


Figure B.27: Experiment 14. Maximum likelihood reconstruction using the Convex algorithm.

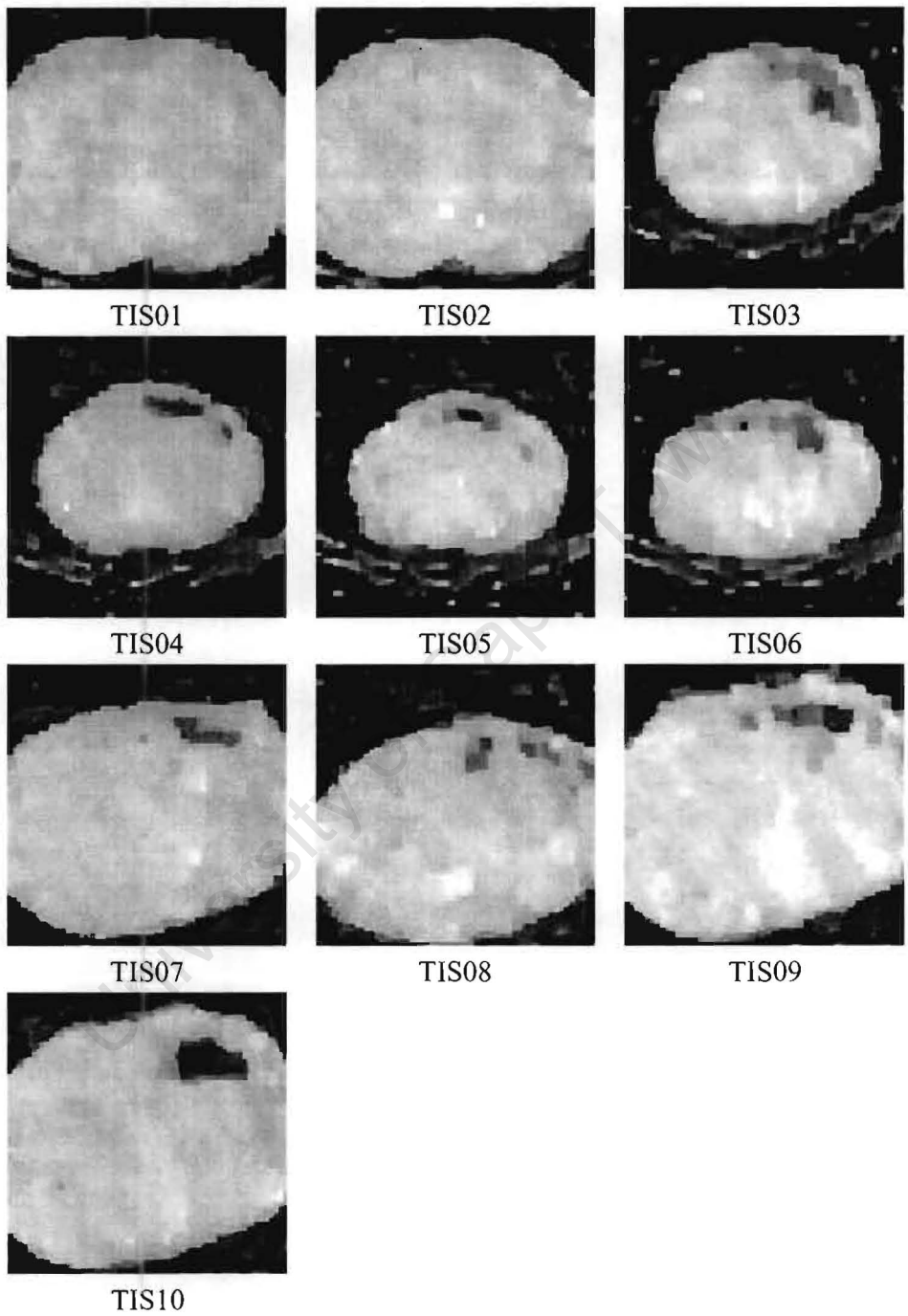


Figure B.28: Experiment 14. MAP reconstruction using the Convex algorithm.

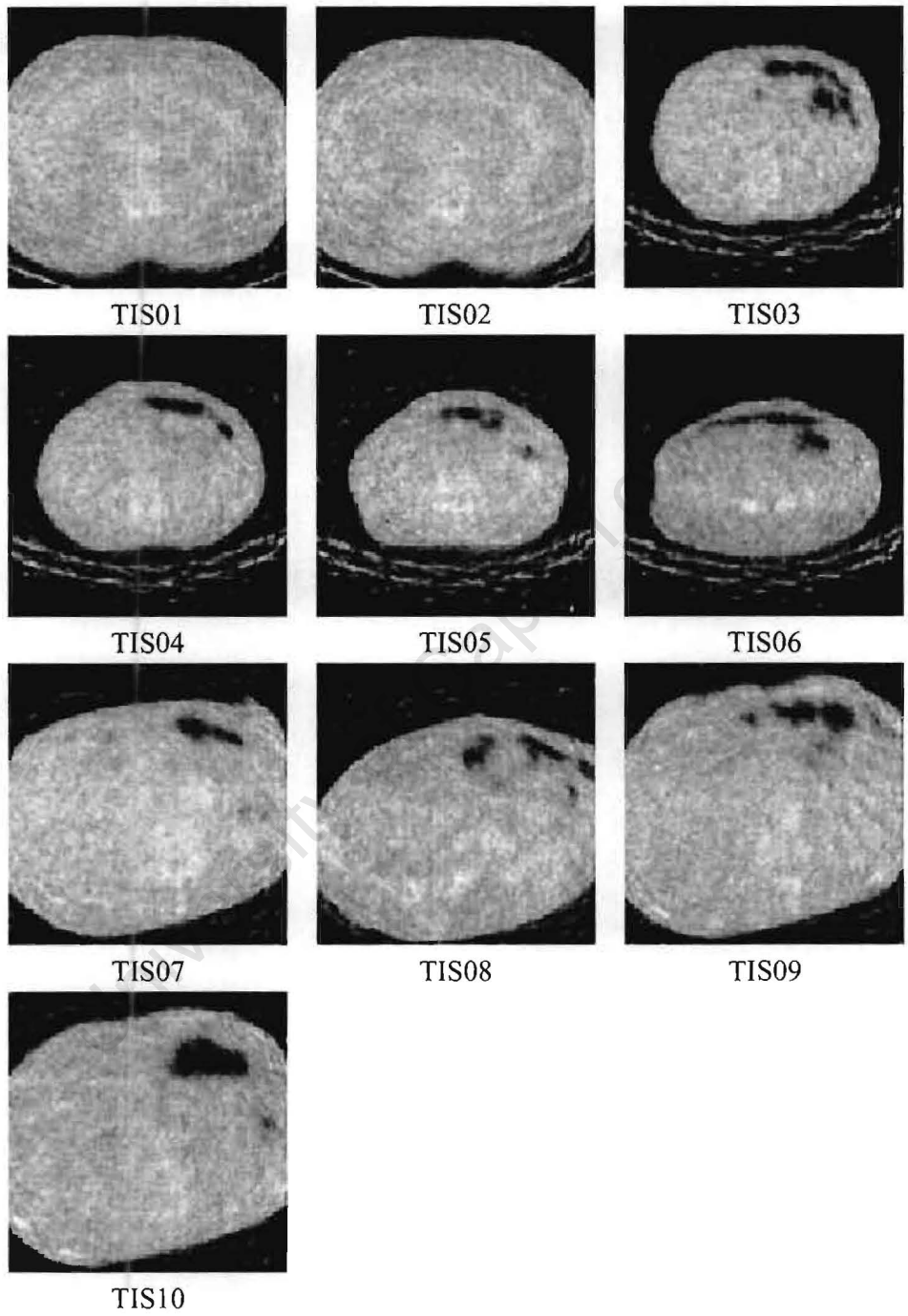


Figure B.29: Experiment 15. Maximum likelihood reconstruction using the Convex algorithm.

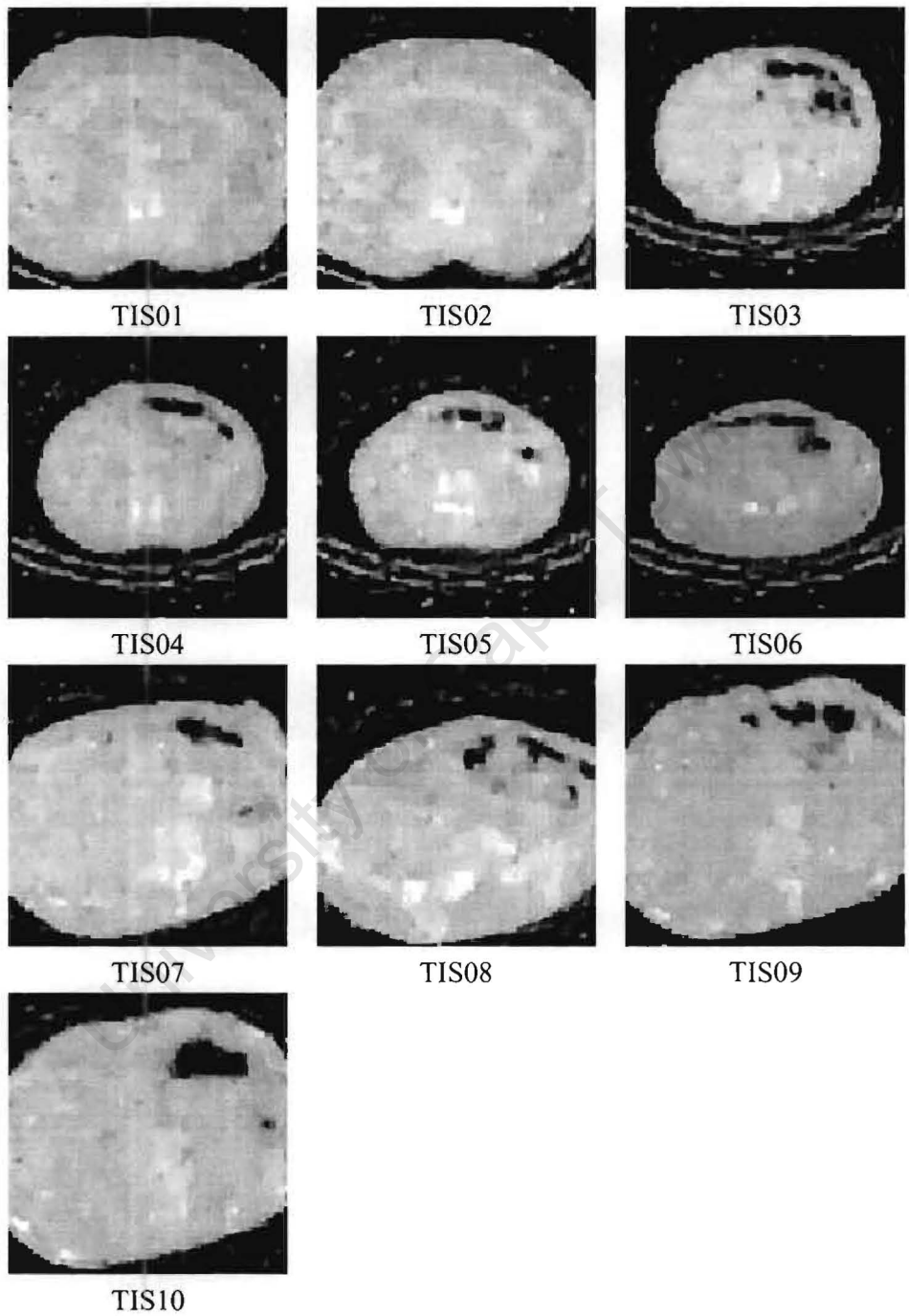


Figure B.30: Experiment 15. MAP reconstruction using the Convex algorithm.

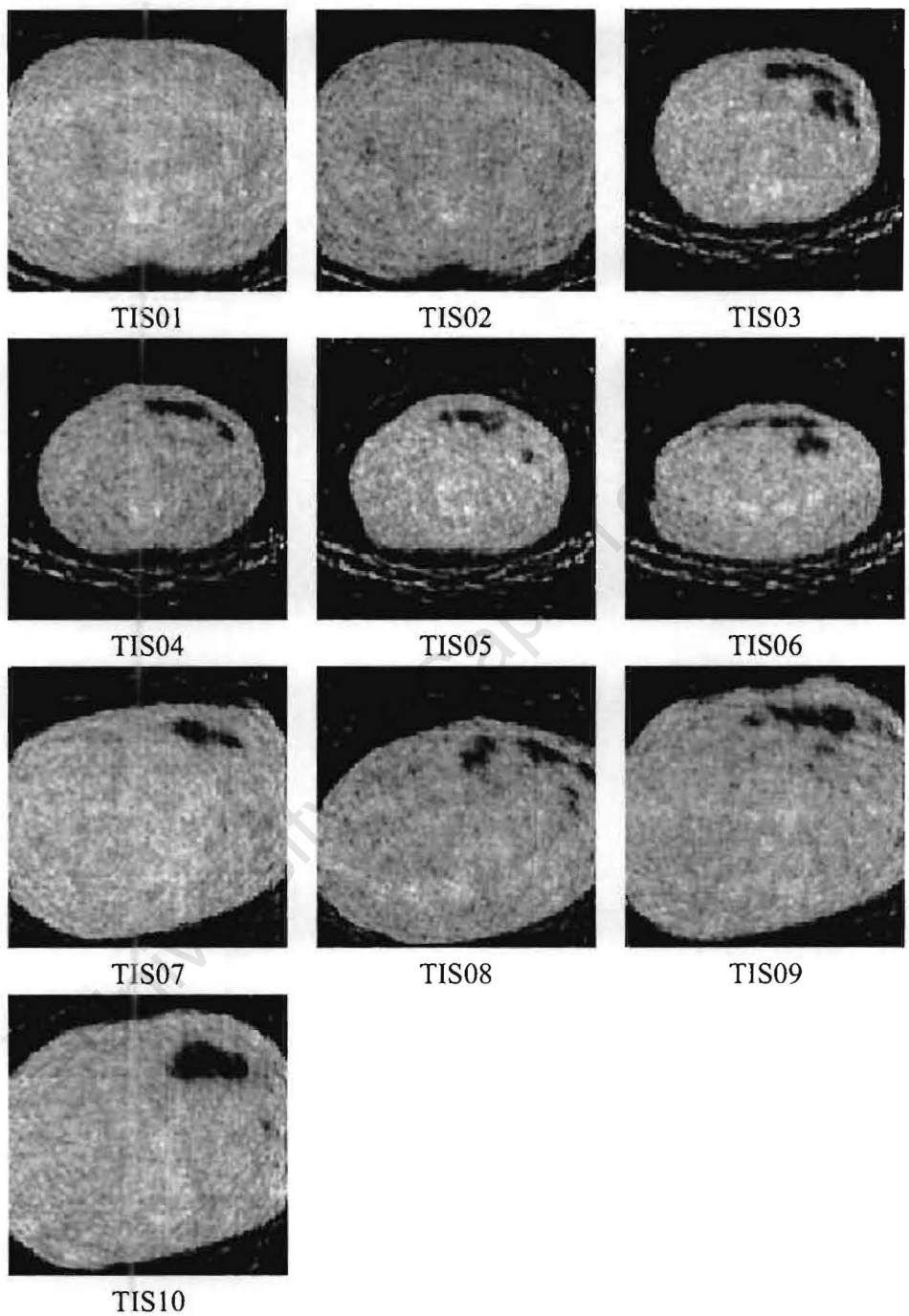


Figure B.31: Experiment 16. Maximum likelihood reconstruction using the Convex algorithm.

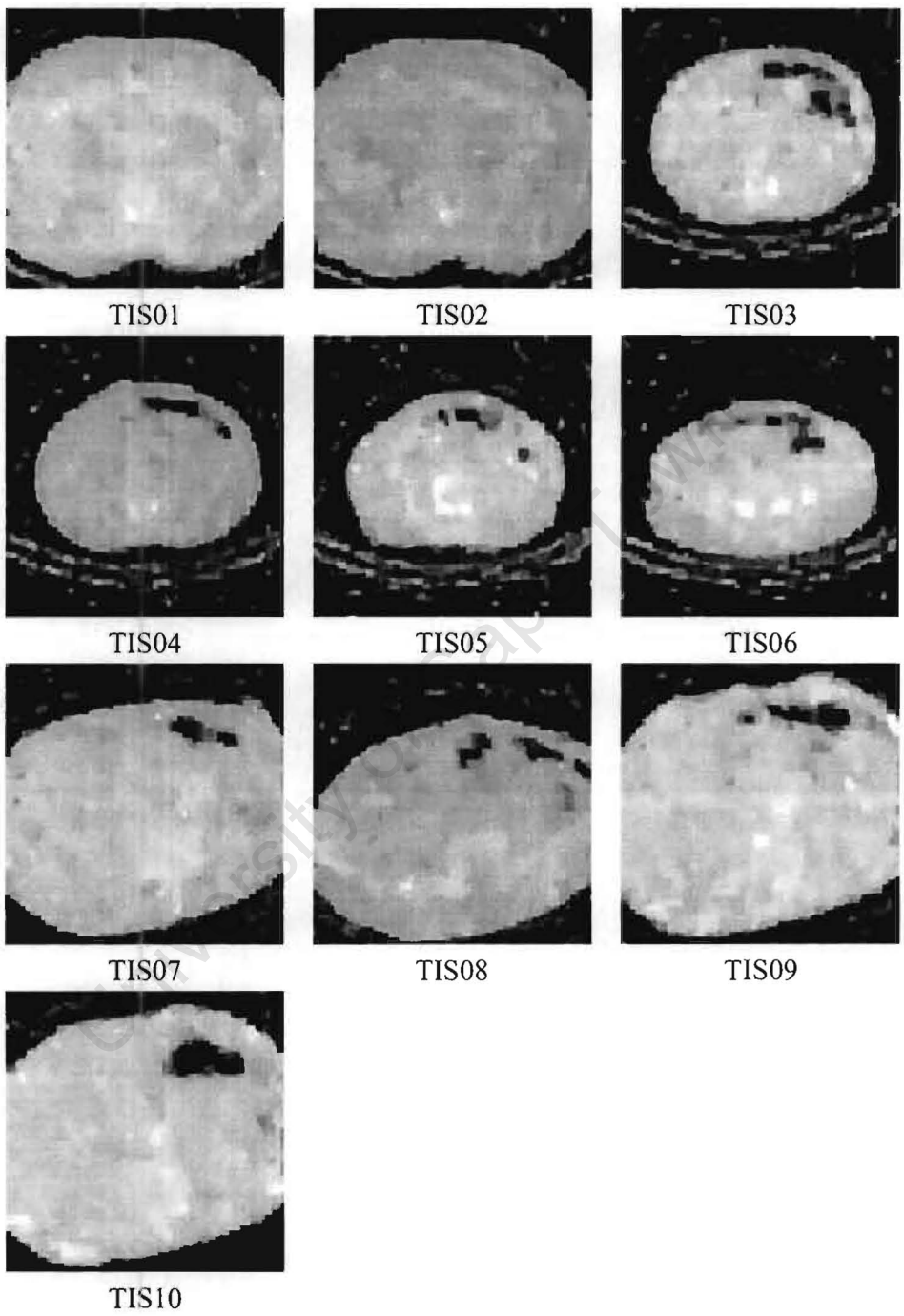


Figure B.32: Experiment 16. MAP reconstruction using the Convex algorithm.

Bibliography

[1] A. Kak and M. Slaney. *Principles of Computerized Tomographic Imaging*. IEEE Press, 1988.

[2] Y. Bresler A.H. Delany. Globally Convergent Edge-Preserving Regularized Reconstruction: An Application to Limited-Angle Tomography. *IEEE Trans. Image Processing*, 7(2):204–221, 1998.

[3] A.P. Dempster, N.M. Laird and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

[4] J. Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *J. R. Statist. Soc. B*, 36(2):192–236, 1974.

[5] J. Besag. Efficiency of pseudolikelihood estimation for simple gaussian fields. *Biometrika*, 64(3):616–618, 1977.

[6] J. Besag. On the Statististical Analysis of Dirty Pictures. *J. R. Statist. Soc. B*, 48(3):259–302, 1986.

[7] J. Besag and P. Green. Spatial Statistics and Bayesian Computation. *J. R. Statist. Soc. B*, 55(1):25–37, 1993.

[8] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[9] C. Bouman and K. Sauer. A Unfied Approach to Statistical Tomography Using Coordinate Descent Optimization. *IEEE Transactions on Image Processing*, 5(3):480–492, March 1996.

[10] B. Buck and V.A. Macaulay, editors. *Maximum entropy in action : a collection of expository essays*. Oxford : Clarendon Press, 1991.

[11] C. Bouman and K. Sauer. A generalized gaussian image model for edge-preserving map estimation. *IEEE Trans. Image Processing*, 2(3):296–310, July 1993.

[12] Y. Censor. Finite Series-Expansion Reconstruction Methods. *Proceedings of the IEEE*, 71(3):409–419, March 1983.

[13] M. DeVilliers. Limited Angle Tomography. Master's thesis, University of Cape Town, 2000.

[14] E. Mumcuoglu, R. Leahy, S. Cherry and Z. Zhou. Fast Gradient-Based Methods for Bayesian Reconstruction of Transmission and Emission PET Images. *IEEE Trans. Med. Imag.*, 13(4):687–701, Dec 1994.

[15] H. Erdougan and J.A. Fessler. Monotonic algorithms for transmission tomography. *IEEE Trans. Med. Imag.*, 18(9):801–814, Sep 1999.

[16] D. Geiger and F. Girosi. Parallel and deterministic algorithms from MRF's: Surface reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5):401–412, 1991.

[17] A. Gelman. Method of moments using Monte Carlo simulation, 1995.

[18] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE PAMI*, 6(6):721, 741 1984.

[19] S. Geman and D. McClure. Bayesian Image Analysis: An application to single photon emission tomography. *Proc. Statistical Computing Section of the American Statistical Association*, pages 12–18, 1985.

[20] J.E. Gentle. *Random Number Generation and Monte Carlo Methods*. Springer, 1998.

[21] C. Geyer. On the Convergence of Monte Carlo Maximum Likelihood Calculations. *Journal of the Royal Statistical Society. Series B*, 56(1):261–274, 1994.

[22] C. Geyer and A. Thompson. Constrained Monte Carlo Maximum Likelihood for Dependent Data. *Journal of the Royal Statistical Society. Series B*, 54(3):657–699, 1992.

[23] G. Gilmore and J. Hemingway. *Practical Gamma-Ray Spectrometry*. Wiley, 1995.

[24] P. J. Green. Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives. *Journal of the Royal Statistical Society. Series B*, 46(2):149–192, 1984.

[25] Grietz. Nobel Presentation Speech. url:http://www.nobel.se/medicine/laureates/1979/presentation-speech.html.

[26] H. Derin and H. Elliott. Modeling and Segmentation of Noisy and Textured Images Using Gibbs Random Fields. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 9(1):39–55, January 1987.

[27] I.A Elbakri and J.A. Fessler. Statistical X-ray computed tomography image reconstruction with beam hardening correction.

[28] J. Besag, P. Green, D. Higdon and K. Mengersen. Bayesian Computation and Stochastic Systems. *Statistical Science*, 10(1):3–41, Feb 1995.

[29] J. Fessler, H. Erdogan and W.B. Wu. Exact Distribution of Edge-Preserving MAP Estimators for Linear Signal Models with Gaussian Measurement Noise. *IEEE transactions on Image Processing*, 9(6):1049–1055, June 2000.

[30] J. Foley, A. van Dam, S. Feiner and J. Hughes. *Computer Graphics:Principles and Practice,Second Edition in C.* Addison-Wesley, 1997.

[31] E. T. Jaynes. Bayesian methods: General background. In J.H. Justice, editor, *Maximum Entropy and Bayesian Methods in Applied Statistics*, pages 1–25, 1985.

[32] C. Ji and L. Seymour. A consistent model selection procedure for markov random fields based on penalized pseudolikelihood. *The Annals Of Applied Probability*, 6(2):423–443, May 1996.

[33] R. Kass and A. Raftery. Bayes Factors. *Journal of American Statistical Association*, 90:773–795, Jun 1995.

[34] S.M. Kay. *Fundamentals of Statistical Signal Processing:Estimation Theory*. Prentice Hall, 1993.

[35] K. Lange and R. Carson. EM Reconstruction Algorithms for Emission and Transmission Tomography. *Journal of Computer Assisted Tomography*, 8(2):306–316, April 1984.

[36] K. Lange and J. Fessler. Globally convergent algorithms for maximum a posteriori transmission tomography. *IEEE Transactions of Image Processing*, 4(10):1430–1438, October 1995.

[37] R.M. Lewitt. Reconstruction Algorithms: Transform Methods. *Proceedings of the IEEE*, 71(3):390–408, March 1983.

[38] Stan Z. Li. *Markov Random Field Modeling in Image Analysis*. Computer Science Workbench. Springer, 2001.

[39] F. Natterer. *The Mathematics of Computerized Tomography*. Wiley, 1986.

[40] J. Ollinger. Maximum-Likelihood Reconstruction of Transmission Images in Emission Computed Tomography via the EM Algorithm. *IEEE Transactions on Medical Imaging*, 13(1):89–101, March 1994.

[41] A. De Pierro. On the relation between the isra and em algorithm for positron emission tomography. *IEEE transactions on Medical Imaging*, 12(2):328–333, June 1993.

[42] J.L. Prince and A.S. Willsky. A geometric projection-space reconstruction algorithm. *Linear Algebra and its Applications*, 130:151–191, 1990.

[43] B. Ripley. *Statistical inference for spatial processes*. Cambridge University Press, 1988.

[44] C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 1999.

[45] S. Kirkpatrick, C.D. Gelatt Jr. and M.P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, May 1983.

[46] K. Sauer and C. Bouman. A Local Update Strategy for Iterative Reconstruction from Projections. *IEEE Transactions on Signal Processing*, 41(2):534–548, Feb 1993.

[47] S.Boyd and L.Vandenberghe. Convex Optimization. url:www.stanford.edu/class/ee364/, Dec 2001.

[48] S.C. Zhu, Y.N. Wu and D. Mumford. Minimax Entropy Principle and Its Application to Texture Modeling. *Neural Computation*, 9(8):1627–1660, 1997.

[49] E.P. Simoncelli and J. Portilla. Texture Characterization via Joint Statistics of Wavelet Coefficient Magnitudes. In *International Conference on Image Processing*. IEEE, 1998.

[50] S.P. Wilson and G. Stefanou. Image segmentation using the double Markov random field, with application to land use estimation. In *ICIP*, pages 742–745. IEEE, 2001.

[51] Y. Dai, E. Rothwell, K. Chen,D. Nyquist. Time-Domain Imaging of Radar Targets Using Sinogram Restoration for Limited-View Reconstruction. *IEEE Transactions on antennas and propagation*, 47(8):1323–1329, Aug 1999.

[52] J. Zhang. The Mean Field Theory in EM procedures for blind Markov Random Field image restoration. *IEEE Trans. Image Processing*, 2(1):27–40, Jan 1993.

[53] D. G. Zill and M. R. Cullen. *Advanced Engineering Mathematics*, chapter Numerical Methods, pages 846–847. PWS, 1992.